BioCOMPUTING

*Technical Report*

# GMAP: A Multi-Purpose Computer Program to Aid Synthetic Gene Design, Cassette Mutagenesis and the Introduction of Potential Restriction Sites into DNA Sequences

**G.P.S. Raghava and Girish Sahni**
Institute of Microbial Technology,
Chandigarh, India

**ABSTRACT**

*A computer program called GMAP has been developed for i) mapping the potential restriction endonuclease (R.E.) sites that can be introduced in a nonambiguous DNA sequence; ii) predicting the mutations required to introduce unique R.E. sites in the nonambiguous DNA sequences; and iii) searching all R.E. sites in ambiguous DNA sequence obtained by reverse translation of a given amino acid sequence. This allows the design of synthetic genes as well as the modular redesign after introducing limited base pair mismatches in wild-type genes in order to adapt them for "cassette" mutagenesis. The GMAP program uses an algorithm based on set theory that reduces the degree of complexity from an exponential to linear function of sequence length. Therefore, the speed of searching for potential R.E. sites in reverse-translated gene sequences and the prediction of new R.E. sites in natural genes by mutations are rapid.*

## INTRODUCTION

An important application of computers in biochemistry is pattern recognition in biological sequences. The need for mapping the restriction endonuclease (R.E.) sites is usually fulfilled by using computers. Finding translationally silent R.E. sites in DNA sequences has become particularly important for biologists, especially those dedicated to the investigation of protein structure/function relationships. The ability to predict potential R.E. sites that are resident in an ambiguous DNA sequence, such as those obtained by reverse translation of protein amino acid sequences, allows one to construct synthetic genes with appropriately placed sites for cutting and joining DNA segment. Similarly, the ability to introduce translationally silent R.E. sites by limited mutagenesis into a nonambiguous DNA sequence (e.g., the open reading frames of natural genes) or in a translationally non-silent manner elsewhere in genes (such as promoters and other control elements that are not expressed into proteins) permits the modular redesign of genes for "cassette" mutagenesis.

The identification of preexisting R.E. sites in DNA sequences is possible with many available programs (4). Handling nonambiguous DNA sequences by available programs is fairly successful because little computational complexity is involved. Several specialized programs are reported that are able to manipulate ambiguous DNA sequences or even protein-coding sequences (1,5,9,11). However, these programs fail to handle protein-coding sequences properly when several ambiguous amino acids are clustered. Presnell and Benner (13) use LISP to represent the protein-coding DNA sequences, but their program is unable to handle long sequences because of the immense complexity involved. For example, since the peptide sequence "Ala Ser Ile" can be represented as GCNAGYATH or GCNTCNATH, these two sequences must be examined separately to determine the placement of all R.E. sites. Thus the ambiguous amino acids (e.g., arginine, leucine and serine), which have six codons, increase the complexity by $2^x$ (where x is the occurrence of ambiguous amino acids in

the amino acid sequence) (13,20). An alternate approach was described (16,17,20) for reducing the complexity of searching for the presence of potential R.E. sites in target amino acid sequences. In this approach, each R.E. recognition sequence is translated into all possible peptides (from three reading frames) instead of reverse-translating protein sequences. These peptides are then searched in the given amino acid sequence, and if any match is found for an R.E. specificity, it indicates that R.E. site is allowed. Although this approach reduces the overall complexity of searching R.E. sites in DNA sequences, it is still complicated because each R.E. recognition sequence, which has dozens of unique peptide patterns, must be separately matched against the target peptide (16,20). Thus, these programs are most applicable when R.E. recognition sequences are short, do not contain any de-

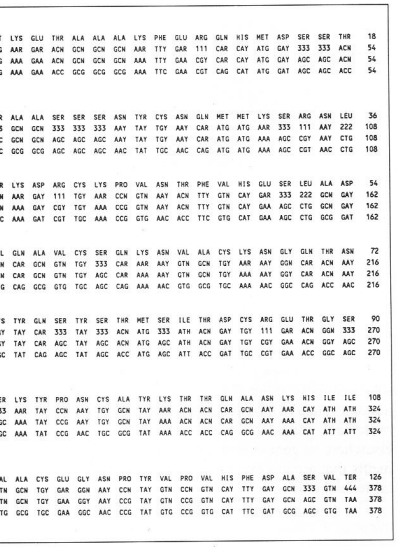


**Figure 1. Reverse translation of the amino acid sequence of bovine pancreatic RNase A into DNA sequences by GMAP.** The codons for the nonambiguous amino acids are represented according to the NC-IUB recommendations (12). However, for ambiguous amino acids, viz., Arg, Leu, Ser, and terminator (Ter), unique numbers 111, 222, 333 and 444 have been employed, respectively, to allow the optimal functioning of the program. 111 is CGN and AGR (Arg), 222 is CTN and TTR (Leu), 333, is AGY and TCN (Ser) and 444 is TAR and TGA (Ter or Stop-codons). An extra Met and Ter residue has been added at the N-terminal and C-terminal end of the known amino acid sequence of mature RNase A (Row A) (2,3) (EMBL: X07283; SWISS-PROT: P00656), to construct functional open reading frames appropriate for synthetic gene design. Three different DNA sequences (rows B–D) have been derived by reverse-translation of the amino acid sequence employing increasingly restrictive codon usage as follows: Row B, fully degenerate using all possible codons; Row C, partially degenerate sequence based on the relatively more frequently utilized codons in *E. coli*; Row D, fully nonambiguous sequence using a single (most frequently used) codon for each amino acid in *E. coli* (19).

generacy and target DNA sequences are relatively small (17). To further lessen the complexity, Jiang et al. (7) described a generic algorithm based on set theory, which reduces the complexity from an exponential to linear function of length of sequence.

The generic algorithm (7) was limited in scope to the search of R.E. sites in ambiguous DNA sequences and/or protein-coded DNA sequences. We have further extended the generic algorithm to incorporate the ability to predict the R.E. sites, which can be introduced in DNA sequence by a limited number (1, 2 or 3 bp per site) of silent mutations. This is a novel capability that is not available in most earlier programs on potential R.E. site prediction. Our algorithm is an extension of the generic algorithm of Jiang et al. (7), hence termed e-generic (e-generic for "extended" generic algorithm). A comprehensive program, GMAP, has been developed based on this algorithm for searching i) the potential R.E. sites in ambiguous DNA sequence and/or protein-coded DNA sequences, and ii) the R.E. sites in nonambiguous DNA sequence and those that can be introduced by 1, 2, or 3 bp site-directed, translationally silent or non-silent mutations. These facets make the program particularly useful both for the redesign of natural genes to incorporate conveniently placed R.E. sites for cassette mutagenesis as well as in synthetic gene design. The program is available free of cost from the EMBL netserver through EMail in both versions compatible for DOS™ and VMS™ operating systems.

## ALGORITHM

We have extended the generic algorithm described earlier (7) for searching R.E. sites in ambiguous DNA sequences and/or protein-coded DNA sequences so that it may also search R.E. sites in nonambiguous DNA sequences that can be introduced by limited site-directed silent (or non-silent) mutations (1, 2 or 3 allowed mismatches per site). Basic elements of DNA sequences such as nucleotides (bases) can be represented as sets, and DNA sequences, whether nonambiguous or ambiguous, can be represented as sequence of these sets. The DNA sequence or recognition sequence of restriction enzyme can be expressed as $x = x_1, x_2, ..., x_r$, where set $x_i$ is $A, B, C, D, G, H, K, M, N, R, S, T, V, W$ or $Y$, and set $A = \{A\}, B = \{G, C, T\}, C = \{C\}, D = \{A, G, T\}, G = \{G\}, H = \{A, C, T\}, K = \{T, G\}, M = \{C, A\}, N = \{A, G, C, T\}, R = \{A, G\}, S = \{C, G\}, T = \{T\}, V = \{A, G, C\}, W = \{A, T\}$ and $Y = \{C, T\}$ (italic capitals stand for sets and normal capitals for nucleotides) (12). Let $x_1$ and $x_2$ be two sets, the set of intersection of both $x_1 \cap x_2$ is a set that contains common elements. If there exists no common element, the result of intersection is an empty set $\phi = \{\}$. In other words, $x_1$ and $x_2$ are said to be matched if $x_1 \cap x_2 \neq \phi$; and mismatched if $x_1 \cap x_2 = \phi$. The potential restriction sites that have one, two or three mismatches in a given DNA sequence can be searched by implementing a set of intersection operations as described below.

Let $\mathbf{x} = x_1, x_2, ..., x_r$ be the recognition sequence of a restriction enzyme and $\mathbf{y} = y_n, y_{n+1}, ..., y_{n+r-1}$ is the segment of natural DNA sequence of length 'r' (equal to the length of

R.E. recognition sequence) from n to n+r-1. Let $TS_i$ denote mismatch index between $x_i$ and $y_{n+i-1}$:

$$TS_i = \begin{cases} 1 \text{ if } x_i \cap y_{(n+i-1)} = \phi \\ 0 \text{ otherwise} \end{cases}$$

Then, the total mismatches may be computed for segment length 'r 'as

$$TS = \sum_{i=1}^{r} TS_i$$

The TS represents the total bp mismatches between the R.E. sequence and DNA segment. There will be a potential R.E. site at position 'n' in DNA sequence if the value of TS is zero, one, two or three (allowed mismatches), which can be introduced in the DNA sequence either without mutation (i.e., a natural site) or by 1-, 2- or 3-bp mutations, respectively. For inspecting whether this site is translationally silent or non-silent, the target DNA sequence is translated to its amino acid sequence, which is then reverse-translated to an ambiguous DNA sequence. Finally, the TS is calculated between the R.E. sequence and the segment of ambiguous DNA sequence from 'n' to 'n+r-1.' A TS value of zero represents a silent potential site; a TS value of 1 or higher signifies a non-silent potential site. Further details of the e-generic algorithm can be obtained by request from the authors.

## IMPLEMENTATION

GMAP was developed on a MicroVax® II under the VMS (version 4.6) operating system (Digital Equipment Corporation, Maynard, MA, USA). It was written in standard PASCAL and compiled with a VAX™ PASCAL (version 3.5) compiler. The code was also compiled under DOS (version 4.01) on an IBM®-compatible PC/AT. It requires no special hardware or graphics to be implemented and runs under VMS or DOS operating systems. The program is interactive, fully menu-driven and allows input or output of data using files. A synthetic RNase A gene (Figure 1) was analyzed by GMAP on an IBM-compatible PC/AT-386 (PLC, Chandigarh, India); using GMAP, it took 63 s CPU time in searching for all the potential sites of 188 type II restriction enzymes (15).

## OPERATION OF THE PROGRAM

The program GMAP is fully menu-driven. Its options and sub-options are shown in Table 1. The 'Input Amino Acid quence' option allows the user to input the amino acid file, and it also allows one to create and update the amino acid sequence file. The sequence data obtained from PIR (Protein Identification Resource) or NBRF (National Biomedical Research Foundation) can also be directly used to create the input amino acid sequence file. The 'Input DNA Sequence' option allows one to create and update the DNA sequence file. The data can be input using the keyboard or from text (or ASCII) file, so that the sequence data extracted from GenBank® or EMBL can be directly used for creating a DNA sequence file. This option also allows one to convert amino acid sequence into DNA sequence by using a user-defined codon preference table. The 'Input R.E. Sequence' option allows the user to create and update the restriction enzyme data file. The prototype restriction endonuclease recognition sequences of type II enzymes (15) are already stored in a file. The 'Input Codon Usage Table' option allows one to create and update the codon preference table. A file containing the codons preferred by *Escherichia coli* (19) is included with the program.

The 'Search R.E. Sites in Amino Acid Sequence' option allows the user to do the following: i) search for all the R.E. sites in fully ambiguous DNA sequence obtained from reverse-translation of amino acid sequence; ii) search the sites for a specific restriction enzyme in reverse-translated ambiguous DNA sequence; iii) reverse translate a given amino

Table 1. Menu of GMAP Computer Program

| Main Menu Options/ Sub-Options | Other Options/Sub-Options |
|---|---|
| **1. Input Amino Acid Sequence** | **Update Sequence File** |
| I. Create amino acid seq. file | I. Display the seq. |
| II. Append amino acid seq. file | II. Insert the seq. |
| III. Update amino acid seq. file | III. Delete the seq. |
| IV. Import data from text file | |
| V. Delete amino acid seq. file | **Search Output (Sites in Amino Acid Seq.)** |
| | I. R.E. sites in order of feed |
| **2. Input DNA Sequence** | II. R.E. sites in order of no. of cuts |
| I. Create DNA seq. file | III. Mapping of all R.E. sites |
| II. Append DNA sequence file | IV. Mapping of specified R.E. sites |
| III. Update DNA sequence file | |
| IV. Import date from text file | **Search Output (Potential Sites in DNA seq.)** |
| V. Input from amino acid seq. | I. Potential R.E. sites in order of feed |
| VI. Delete DNA sequence file | II. Potential sites not present in natural seq. |
| | III. Mapping of potential R.E. sites |
| **3. Input R.E. Sequence** | IV. Mapping of potential sites absent in natural seq. |
| I. Create R.E. sequence file | |
| II. Append R.E. sequence file | **Search Ouptut (Natural Sites in DNA seq.)** |
| III. Update R.E. sequence file | I. Natural R.E. sites in order of feed |
| IV. Import data from text file | II. Natural R.E. sites in order of no. of cuts |
| V. Delete R.E. sequence file | III. Mapping of all natural R.E. sites |
| | IV. Mapping of potential sites absent in natural seq. |
| **4. Input Codon Usage Table** | |
| I. Help about codon usage table | |
| II. Create codon usage table file | |
| III. Update codon usage table file | |
| IV. Delete condon usage table file | |
| **5. Search R.E. Sites in Amino Acid Sequence** | |
| I. R.E. sites in ambiguous DNA seq. | |
| II. Specified sites in ambig. DNA seq. | |
| III. Reverse translate into DNA sequence | |
| IV. R.E. Sites in partially ambig. DNA seq. | |
| V. Specified sites in partially ambig. DNA seq. | |
| **6. Search R.E. Sites in DNA Sequence** | |
| I. All potential R.E. sites in DNA seq. | |
| II. Specific potential sites in DNA seq. | |
| III. Translate DNA seq. into amino acid seq. | |
| IV. All (natural) R.E. sites in DNA seq. | |
| V. Specified (natural) R.E. sites in DNA seq. | |
| **7. Output DNA/Amino Acid/R.E./Codon Usage Table** | |
| I. DNA sequence | |
| II. Amino acid sequence | |
| III. Restriction enzyme and recognition seq. | |
| IV. Codon usage table | |

acid sequence into fully or partially ambiguous DNA sequence or into completely nonambiguous DNA sequence using user-defined codon preference; iv) search all R.E. sites in partial (or nonambiguous) DNA sequence obtained from reverse translation of amino acid sequence employing the user-defined codon preference table; and v) search the sites for user-specified enzyme in partially ambiguous or completely nonambiguous DNA sequence obtained from reverse-translation of amino acid sequence with user-defined codon usage.

The 'Search R.E. Sites in DNA Sequences' option allows the user to do the following: i) search all the potential R.E. sites that can be introduced in DNA sequence by limited site-directed silent/non-silent mutagenesis and the number of mutations required to introduce a site; ii) search the potential sites for a specific restriction enzyme, which can be introduced in DNA sequence by site-directed silent/non-silent mutagenesis, and the number of mutations required to introduce a site; iii) translate the DNA sequence into amino acid sequence; iv) search the preexisting sites of all R.E.s in DNA sequence; and v) search existing sites of a specific R.E. in the DNA sequence.

The 'Output DNA/Amino Acid/R.E./Codon Usage Table' option allows the display (or printout or save in file) of the amino acid sequence, DNA sequence, restriction enzyme data and codon preference usage table. Besides the main options and sub-options, there are other options that allow the user to output the results in the desired format (Table 1). The program has the ability to cope with both palindromic and non-palindromic recognition sites.

## RESULTS AND DISCUSSION

Protein engineering by genetic means is currently one of the foremost techniques of studying the relationship between structure and function of a protein. Recombinant DNA technology, particularly the use of the relatively straightforward polymerase chain reaction (PCR) methods, allows the facile incorporation of site-specific alterations in amino acid sequence by modifying the target DNA (6,18). This approach is greatly facilitated if a given (wild-type) gene is so altered (by prior, limited mutagenesis) as to allow the ready replacement of a DNA segment (cassette) of the gene with another synthetic or PCR-generated segment that codes for the desired alteration in the protein sequence. The alternative to this semi-synthetic approach is the complete redesign of genes by DNA synthesis methods. Here, suitable cassettes can be pre-designed at will by introducing appropriately placed translationally silent R.E. sites into the sequence (8,13,20). In this case, the "sequence space" can be further limited by restricting the usage of the codons for different amino acids based on the relative frequencies of their use in the host organism employed for the expression of the designed gene (19). Thus, the prediction of silent R.E. sites in target sequences is of great importance in protein engineering projects.

The GMAP program is suitable both for the design of totally synthetic genes based on protein sequence (ambiguous DNA sequence) and for the redesign of natural genes, with nonambiguous DNA sequences, by limited site-specific mutagenesis in order to obtain a modular cassette arrangement. The program has been successfully implemented for achieving both the objectives by using two examples: namely, RNase A and streptokinase. One particular advantage that "custom-tailored," synthesis-based gene design enjoys over the manipulation of natural genes is the possibility of altering the codon preferences (19) for different amino acids in the protein in consonance with the requirements of various host-cell protein synthesis machineries. In this case too, GMAP offers the option of restricting codon preferences to user-defined dictates. Thus, instead of a totally non-preferential usage of the respective codons for different amino acids, a partially ambiguous DNA sequence is generated by reverse translation of the amino acid sequence according to specified codon usage, which is then analyzed for mapping unique potential R.E. sites that are translationally silent. The user could then incorporate any or all of these sites into the final DNA sequence chosen for synthesis.

The applicability of GMAP to *de novo* gene design has been tested using the example of RNase A (Figures 1 and 2). The amino acid sequence of RNase A is well known, and this protein has served as a favorite model system for numerous protein structure-function studies over the years (2,3). The amino acid sequence of RNase A was reverse-translated by GMAP into DNA sequences with varying degrees of ambiguity due to different codon usage (Figure 1). The maps of unique R.E. sites in either the totally nonambiguous DNA sequence (choosing only a single, most frequently used, *E. coli* codon for each amino acid) or partially ambiguous DNA sequence (using only the relatively more highly used codons of *E. coli*) or totally ambiguous DNA sequence (choosing all degenerate codons) were then determined (Figure 2). The example of RNase A clearly illustrates the successful applicability of GMAP for predicting the useful R.E. sites to be designed in the different regions of a gene with different codon preference constraints during its *de novo* synthesis. Despite a relative limitation on the degree of freedom of sequence choice due to restriction on codon usage, an adequate number of unique sites is still available in RNase A in the case of the partially and fully degenerate DNA sequences that permit the design of a gene with a cassette arrangement useful for its (future) mutagenenic manipulation. In cases where this may not be possible, one has the option of exploring double-, triple- and higher order-cutter R.E. sites for possible manipulation. In such cases, potential site(s) located in the area of interest can be retained while sites elsewhere for the same enzyme can be simply abolished (in a translationally silent manner) if feasible on the basis of sequence degeneracy.

Although gene design by total *de novo* synthesis is a powerful tool for protein engineering, a convenient, albeit somewhat less powerful, approach is to investigate natural genes for the possibility of introducing new R.E. sites into the nonambiguous DNA sequence through a limited number of base pair mismatches. Since most of the cloned genes are of natural origin rather than chemically synthesized *in vitro*, the ability to manipulate wild-type genes for protein engineering purposes is of special interest to molecular biologists. Apart from conferring advantages during modular mutagenesis, the

placement of unique R.E. site(s) (RFLPs) near or within a target sequence to mark that gene or mutagenesis event also provides a useful tool in genetic experiments. This feature is clearly illustrated in the example of streptokinase, an important thrombolytic protein (Figure 3). The known DNA sequence of streptokinase (10) was first analyzed by GMAP for naturally present, single-cutter (i.e., unique) R.E. sites (column 1 in Figure 3). Although several sites are placed uniformly throughout most of the gene (a key consideration in obtaining a modular arrangement of cassettes in genes), several segments (indicated by boxes, marked A, B and C in column 1) lack such sites. However, upon limited mismatching

**PREDICTED SINGLE CUTTER R.E. SITES IN RNASE A GENE**

| Fully Ambiguous Sequence | | Partially Ambiguous Sequence | | Non-Ambiguous Sequence | |
|---|---|---|---|---|---|
| Postion (Nucl. No.) | Restriction Enzyme | Postion (Nucl. No.) | Restriction Enzyme | Postion (Nucl. No.) | Restriction Enzyme |
| 13 | NotI | 11 | AlwNI | 11 | DsaI/NspBII/SacII |
| 28 | BsrBI | 13 | NotI | 22 | ApoI |
| 33 | PflMI | 22 | ApoI | 23 | TspEI |
| 44 | BaeI | 25 | AsuII | 25 | AsuII |
| 48 | AccI/SalI | 33 | PflMI | 37 | NdeI |
| 52 | SpeI | 37 | NdeI | 71 | ApaBI |
| 55 | Eco47III | 51 | BsiI | 101 | MaeIII |
| 71 | ApaBI | 55 | HaeII/Eco47III | 115 | DpnI/MboI |
| 97 | XbaI | 71 | ApaBI | 126 | BetI/AgeI |
| 98 | FauI | 115 | MboI/DpnI | 129 | HphI |
| 105 | EcoNI | 130 | HpaI | 132 | XmnI |
| 116 | ClaI | 131 | MseI | 156 | EciI |
| 130 | HpaI | 132 | XmnI | 159 | FokI |
| 132 | XmnI | 134 | AflIII | 204 | CfrI |
| 144 | BspHI | 144 | BspHI | 205 | HaeIII |
| 150 | HindIII | 156 | EciI | 226 | AluI |
| 152 | EspI | 159 | FokI | 228 | SfeI |
| 187 | AclI | 160 | AatII | 250 | BsrDI |
| 204 | BalI | 163 | BspGI | 300 | TaqII(2) |
| 237 | BsaBI | 187 | AclI | 312 | Tth111II |
| 242 | BccI | 204 | BalI | 364 | SfaNI |
| 264 | DrdII | 213 | MfeI | | |
| 265 | DraII/PpuMI/XhoII/BamHI | 237 | BsaBI | | |
| 286 | MstI | 255 | NruI | | |
| 300 | TaqII(2) | 262 | BsrI | | |
| 337 | BstEII | 286 | MstI | | |
| 344 | SplI | 300 | PleI/HinfI/MlyI/TaqII(2) | | |
| 346 | SnaBI | 302 | DdeI/Bce831 | | |
| 365 | MluI/AvaIII | 323 | Pfl1108I | | |
| | | 337 | BstEII | | |
| | | 344 | SplI | | |
| | | 346 | SnaBI/BsaAI | | |
| | | 349 | FinI | | |
| | | 351 | CauII | | |
| | | 354 | AvaII/AsuI | | |
| | | 364 | DrdI | | |

**Figure 2. Prediction of R.E. sites in RNase A gene sequences obtained by reverse translation of the amino acid sequence.** The different DNA sequences shown in Figure 1 were used to map the R.E. sites with GMAP. Only the single-cutter (unique) sites are shown for simplicity. The full ensemble of all possible translationally silent R.E. sites for all R.E. recognition specificities in the ambiguous sequence can also be obtained in a single output (data not shown). Note also, that the use of a highly restrictive codon usage in the case of the nonambiguous DNA sequence (first column from right) yields a large segment (nucleotides 228 to 312) without many intervening R.E. sites. However, relaxing the codon usage to the relatively more frequently used codons in *E. coli* (middle column), or to a fully ambiguous DNA sequence using a totally non-preferential *E. coli* codon usage (first column from left), results in the appearance of several new unique R.E. sites in the DNA segment.

(i.e., 1, 2 or 3 bp per R.E. site), several new, unique sites are seen to appear (see corresponding boxes in column 2 in Figure 3). The choice is even further extended when new double- and triple-cutter enzyme sites (3rd and 4th columns, respectively) are explored. In this case, the user has the option of scrutinizing the potential double-, triple- or quadruple-cutter sites permissible in the sequence after limited mismatching and then to allow only those sites that permit the placement of a unique site in the area of interest while leaving other sites intact. Note that only those sites that are generated upon 1-bp mismatching have been shown in Figure 3 (because they turned out to be adequate for the example chosen). In

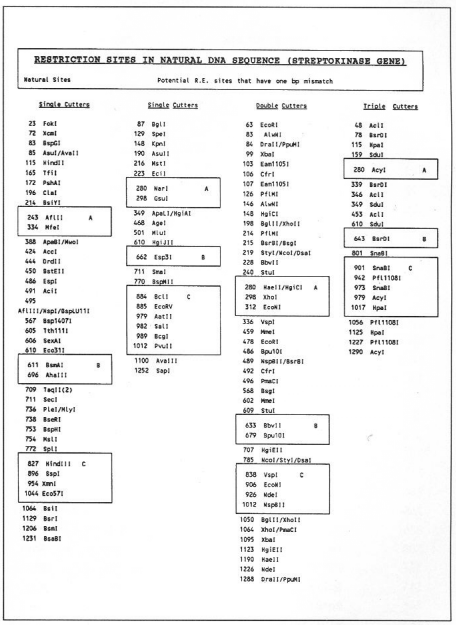


**Figure 3. Prediction of potential R.E. sites in a natural gene streptokinase with GMAP.** The nucleotide sequence of streptokinase (Reference 10) (EMBL: K02986 and X13400; SWISS-PROT: P00779 and P10519) was used to determine the preexisting R.E. sites present in the gene (only single-cutters are shown). The new, translationally silent R.E. sites (single-, double- or triple-cutters) that can be generated after allowing limited mismatches (1, 2 or 3 bp per site) were determined by GMAP. Only the sites obtained by single bp mismatch are shown for the sake of simplicity. Those segments in the natural streptokinase gene that are deficient in preexisting R.E. sites are highlighted by boxes A, B and C. The additional potential R.E. sites generated in the corresponding region of the gene after 1-bp mismatch are also boxed (note that the "new" R.E. sites predicted by GMAP are such that they are not present in the natural DNA sequence). Thus, additional unique (or nonunique) R.E. sites can be readily chosen from the array of available single-, double- or triple-cutter sites after limited mismatching and introduced by oligonucleotide-directed, site-specific mutagenesis to obtain a derivative of the natural gene with the desired cassette arrangement.

cases where this is not adequate, the ensemble of potential R.E. sites permissible through 2, 3 or a higher order of bp mismatching will likely be sufficiently vast to permit the introduction of useful sites virtually in any region of a gene. Alternatively, if one wishes to incorporate potential R.E. sites into DNA that can be generated only by altering the encoded amino acid sequence, GMAP offers this option to the user. This scenario is particularly useful when the constraint of "translation silence" is not needed for the mutagenesis of regions other than the open reading frames, such as the control elements of genes. A pertinent example of this type of application is when enhancing the expression of whole genes by cassette mutagenesis wherein one desires to cut just outside of a coding region in order to fuse it to a stronger promoter.

## AVAILABILITY

GMAP is freely available either by request from authors or from the EMBL (source code and compiled programs for VAX/VMS and DOS computers) through (EMail) and anonymous file transfer (ftp). EMail can be sent to the Internet address of EMBL (14), netserv@embl-heidelberg.de, by typing the following commands (only one per line in the body of the message):

HELP DOS_SOFTWARE
GET DOS_SOFTWARE:GMAP.UUE **OR**
HELP VAX_SOFTWARE
GET VAX_SOFTWARE:GMAP.UUE

This provides the programs in a uu-encoded form, which can be processed according to the information given in the HELP files. Alternatively, fully functional programs can be obtained through ftp from 'ftp.embl-heidelberge.de' (192.54.41.33) by giving the username ANONYMOUS and the user's EMail address as the password. The DOS version of GMAP program can be obtained by switching the directory (command CD PUB/SOFTWARE/DOS), looking for available files (command DIR), altering the transfer mode to binary (command BINARY), and ordering the desired program (command GET GMAP$.EXE). After transfer, the ftp session is terminated (command QUIT). GMAP$.EXE is a self-unstuffing DOS application, providing the interested user with the program, all necessary files and a read-me document.

## ACKNOWLEDGMENTS

This is communication No. 019/93 from the Institute of Microbial Technology, Chandigarh, India, and was supported by grants from the Council of Scientific and Industrial Research and the Department of Biotechnology, Government of India. The authors are grateful for the suggestions offered by the anonymous referees. The authors are also thankful to Dr. Grish C. Varshney and Mr. Mahavir Yadav for their help in preparing the manuscript. Internet: raghava@imtech.ernet.in

## REFERENCES

1. **Arentzen, R. and W.C. Ripka.** 1984. Introduction of restriction enzyme sites in protein-coding DNA sequences by site-specific mutagenesis not affecting the amino acid sequence: a computer program. Nucleic Acids Res. *12*:777-787.
2. **Beintema, J.J., C. Schueller, M. Irie and A. Carsajna.** 1988. Molecular evolution of the ribonuclease superfamily. Prog. Biophys. Mol. Biol. *51*:165-192.
3. **Blackburn, P. and S. Moore.** 1982. Pancreatic ribonuclease, p. 317-329. *In* P.D. Boyer (Ed.), The Enzymes, Vol. 15. Academic Press, New York.
4. **Cannon, G.** 1990. Nucleic acid sequence analysis software for microcomputers. Anal. Biochem. *190*:147-153.
5. **de Boer, J.G.** 1991. MARS: A program to find potential restriction sites. Comput. Appl. Biosci. *7*:267.
6. **Higuchi, R., B. Krummel and R.K. Saiki.** 1988. A general method of *in vitro* preparation and specific mutagenesis of DNA fragments: study of protein and DNA interactions. Nucleic Acids Res. *16*:7351-7367.
7. **Jiang, K., J. Zheng and S.B. Higgins.** 1991. A generic algorithm for finding restriction sites within DNA sequences. Comput. Appl. Biosci. *7*:249-256.
8. **Libertini, G. and A. Di Donato.** 1992. Computer-aided gene design. Protein Eng. *5*:821-825.
9. **Little, J.W. and D.W. Mount.** 1984. Creating new restriction sites by silent changes in coding sequences. Gene *32*:67-73.
10. **Malke, H., B. Roe and J.J. Ferretti.** 1985. Nucleotide sequence of the streptokinase gene from *Streptococcus equisimilis* H 46A. Gene *34*:357-362.
11. **Mount, D.W. and B. Conrad.** 1986. Improved programs for DNA and protein sequence analysis on the IBM personal computer and other standard computer systems. Nucleic Acids Res. *14*:443-454.
12. **NC-IUB recomendations (Nomenclature Committee of the International Union of Biochemistry).** 1985. Nomenclature for incompletely specified bases in nucleic acid sequences. Eur. J. Biochem. *150*:1-5.
13. **Presnell, S.R. and S.A. Benner.** 1988. The design of synthetic genes. Nucleic Acids Res. *16*:1693-1702.
14. **Rice, C.M., R. Fuchs, D.G. Higgins, P.J. Stoehr and G.N. Cameron.** 1993. The EMBL data library. Nucleic Acids Res. *21*:2967-2971.
15. **Roberts, J.R. and D. Macelis.** 1993. REBASE-restriction enzymes and methylases. Nucleic Acids Res. *21*:3125-3137.
16. **Shankarappa, B., D.A. Sirko and G.D. Ehrlich.** 1992. A general method for the identification of regions suitable for site-directed silent mutagenesis. BioTechniques *12*:382-384.
17. **Shankarappa, B., K. Vijayananda and G.D. Ehrlich.** 1992. SILMUT: A computer program for the identification of regions suitable for silent mutagenesis to introduce restriction enzyme recognition sequences. BioTechniques *12*:882-884.
18. **Vallette, F., E. Mege, A. Reiss and M. Adesnik.** 1989. Construction of mutant and chimeric genes using the polymerase chain reaction. Nucleic Acids Res. *17*:723-733.
19. **Wada, K., Y. Wada, F. Ishibashi, T. Gojobori and T. Ikemura.** 1992. Codon usage tabulated from the GenBank genetic sequence data. Nucleic Acids Res. *20*:2111-2118 .
20. **Weiner, M.P., and H.A. Scheraga.** 1989. A set of Macintosh computer programs for the design and analysis of synthetic genes. Comput. Appl. Biosci. *5*:191-198.

**Address correspondence to:**

G.P.S. Raghava
*IMTECH*
*Sector 39A*
*Chandigarh 160 014, India*