Bull. Inst. Chem. Res., Kyoto Univ., Vol.71, No.3, 1993

COMMUNICATION

Nucleotide Sequence of the Gene of *Hga*I Restriction Endonuclease

Hiroyuki SUGISAKI*

Received July 20, 1993

The nucleotide sequence reported in this paper has been submitted to the DDBJ, EMBL, and GenBank Data Bank with accession number D17388.

The abbreviations used are: bp, base pair(s); kb, kilobase pair(s); ORF, open reading frame; R-M system, restriction-modification system; R*Hga*I, *Hga*I restriction endonuclease; M*Hga*I, *Hga*I modification methylase; IPTG, isopropyl-β-D-thiogalactopyranoside

KEY WORDS: DNA cloning / *Haemophilus gallinarum* / Type IIS restriction-modification system / Expression in *E.coli*

The genes of a variety of the type II restriction-modification (R-M) systems that recognize symmetrical DNA sequences have been cloned and their gene organization and nucleotide sequences have been analyzed [see the reviews[1-3]]. As concerns the genes consisting of the type IIS R-M systems that recognize asymmetrical DNA sequences, however, only those of the *Fok*I[4], *Mbo*II[5] and *Sts*I[6] [the endonuclease recognizes the same DNA sequence as *Fok*I endonuclease but cleaves at different positions[7]] R-M systems have been reported so far.

The *Hga*I R-M system found in *Haemophilus gallinarum*, that belongs to this category, recongnizes a 5 bp segment of DNA, consisting of 5'-GACGC-3' in one strand and 3'-CTGCG-5' in the other (*Hga*I recognition site). *Hga*I restriction endonuclease (R*Hga*I) introduces staggered cleavages 5 and 10 nucleotides downstream from the recognition sequences[8,9]. Experiments have been designed to clone the genes consisting of the system and a DNA fragment carrying the genes of two *Hga*I modification methylase (M*Hga*I) has been successfully isolated from a genomic DNA library of *H.gallinarum*[10]. The two methylase genes are respectively responsible for modification of the internal cytosine residues of the CTGCG and GACGC strands and render DNA resistant to digestion by R*Hga*I. However, the R*Hga*I gene was not identified. Since the genes of restriction endonucleases are usually linked to those of the corresponding methylases on chromosome, truncated open reading frame 3 (ORF3) previously identified, which is located just downstream from the two M*Hga*I genes, is most likely to be the corresponding gene (see Fig.2 in Ref.10). The separation of the domains for DNA sequence recognition and cleavage makes type IIS restriction endonucleases interesting objects for studies of DNA-protein recognition. Therefore, an experiment was designed to clone the complete gene from the chromosomal DNA.

As to maintain the plasmid carrying a restriction endonuclease gene in *E.coli* cells, coexistence of the corresponding methylase gene is indispensable. First, the plasmid carrying the

*椙崎弘幸: Research Facility of Nucleic Acids, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611, Japan

two M*Hga*I genes was constructed. The 2.9 kb region of pKS318[10] (from the leftmost *Hin*dIII site to the *Xba*I site at nucleotide position 2909, see Fig.2) was inserted into the *Hin*dIII-*Xba*I region in the multicloning sites of pUC19[11]. Deletion was introduced from the *Xba*I site and a clone carrying a DNA fragment of nucleotide positions 1 to 2497 was isolated. The coding region for the two M*Hga*I genes could be just genetated as a 2.4 kb fragment from the plasmid construct by R*Nsp*(7524)V-R*Eco*RI digestion and inserted into the *Eco*RV site of pACYC184[12] by blunt end ligation. The resulting construct carrying the two M*Hga*I genes under control of the *tet* promoter was named pKS430 (see Fig.2) and introduced into *E.coli* MCH1061 cells[10].

Next, a DNA fragment adjacent to the insert of pK318 on the *H.gallinarum* chromosome was searched. When digests of the chromosomal DNA by various restriction endonucleases were subjected to the Southern blotting hybridization[13] using, as probe, the 390 bp *Xba*I-*Hin*dIII fragment (nucleotide positions 2,909 to 3,399, see Fig.2) which is located in the coding region for ORF3, an approximately 4.5 kb *Cla*I fragment was identified (data not shown). The DNA fragment would be contain the region adjacent at the right side to the insert and a sufficient coding capacity for complete ORF3 because the *Cla*I site is located just upstream from the gene (see Fig.2). Complete *Cla*I digests of the chromosomal DNA were fractionated by electrophoresis on 0.7% agarose gel, and fractions predominantly containing the fragments of 3.5-5.5 kb were collected. The sized *Cla*I fragments were ligated to a phosphatase-treated *Acc*I digest of pUC19 and introduced into *E.coli* MCH1061 carrying pKS430. Among the recombinants examined by colony hybridization[13] using the *Xba*I-*Hin*dIII fragment as probe, one carrying ORF3 under control of the *lac* promoter in a single 4.5 kb *Cla*I fragment was named pKS456. The R*Hga*I activity was actually detected in the extract from the *E.coli* cells by induction with isopropyl- β -D-thiogalactopyranoside (IPTG). I, therefore, concluded that the 4.5 kb *Cla*I fragment contains the R*Hga*I gene.

The nucleotide sequence of the region adjacent to the insert of pKS318 was determined using the dideoxy sequencing method[14,15]. The DNA sequence of the 1,8-kb leftmost *Cla*I to *Hin*cII region of pKS456 is shown in Fig.1. Analysis of coding capacities indicated that ORF3 terminates at nucleotide position 3808.

To verify that ORF3 is indeed the R*Hga*I gene, the endonuclease activity was determined by the *in vitro* assay procedure. The leftmost 1.6 kb region of the 4.5 kb *Cla*I fragment (from the *Bam*HI site on the pUC19 moiety to the *Eco*T22I site, see Fig.2) which encompasses entire ORF3 was inserted into the *Bam*HI-*Pst*I region in the multicloning sites of the expression vector pKK223-3 (Pharmacia P-L Biochemicals Inc.). The resulting construct carrying ORF3 under control of the *tac* promoter was named pKS475. The plasmid construct was introduced into *E.coli* MCH1061 cells carrying pKS430 and after induction with IPTG, protein expressed from the gene was purified by phosphocellulose (Whatman), heparin-agarose (Bethesda Research Laboratories), and hydroxyapatite (Clarkson Chemical Co.) columns. When the various DNAs were digested with the purified protein fraction and authentic R*Hga*I isolated from *H.gallinarum* cells, the same restriction patterns were obtained (data not shown). The results clearly indicated that ORF3 was the R*Hga*I gene. In ORF3, the ATG codon occurred at nucleotide positions 2345 and 2360. The predicted molecular weights of proteins beginning with ATG at nucleotide positions 2345 and 2360 were 56,734 and 56,119, respectively. I assigned the ATG codon at 2360 to the initiation site for translation of the endonuclease for only this one was accompanied by a putative ribosomal-binding sequence, 5'-GGAG-3',

```
                                                CGATGAAAATATAGAATTTCCACCAGGCTCCAGTGAGACT  2280
TATATTAGAACTATTATTGGTGAGGCTATTCCACCCAAATTACTTTCAGCAATATGTTTT  2340
CCTGATGGAGAAAATATTAATGTTAAATGATGATCAAATTTGGATATTTAAAAAACATAC  2400
      Met  Glu  Lys  Ile  Leu  Met  Leu  Asn  Asp  Asp  Gln  Ile  Trp  Ile  Phe  Lys  Lys  His  Thr    19
TAATAATATTCAGCTATTAATTGAAGTTGCTTTGTATTTAAAAAGCAATAAATCTTCTGT  2460
  Asn  Asn  Ile  Gln  Leu  Leu  Ile  Glu  Val  Ala  Leu  Tyr  Leu  Lys  Ser  Asn  Lys  Ser  Ser  Val    39
TTCAAAAAAAGATAAAGATGCGATGTATGATATATTCAGTGAATCTGAATTATATAATCC  2520
  Ser  Lys  Lys  Asp  Lys  Asp  Ala  Met  Tyr  Asp  Ile  Phe  Ser  Glu  Ser  Glu  Leu  Tyr  Asn  Pro    59
AAGAGAATCTTTACGTGATAAACCTTTGGACACAATTAATCATAAATTAGATGGGTTATC  2580
  Arg  Glu  Ser  Leu  Arg  Asp  Lys  Pro  Leu  Asp  Thr  Ile  Asn  His  Lys  Leu  Asp  Gly  Leu  Ser    79
TTATTTTATGTTTGGGTACTCAGATAGAATAAATGATGAAAATAAATTTATTTTTAGTCC  2640
  Tyr  Phe  Met  Phe  Gly  Tyr  Ser  Asp  Arg  Ile  Asn  Asp  Glu  Asn  Lys  Phe  Ile  Phe  Ser  Pro    99
ATTAGGAAATCTTTTCCTGAAATATCTACATGATAAAGATAAATTAAGCAAAATTTTTTC  2700
  Leu  Gly  Asn  Leu  Phe  Leu  Lys  Tyr  Leu  His  Asp  Lys  Asp  Lys  Leu  Ser  Lys  Ile  Phe  Ser   119
TTGTATGCTTATTAGTATGCAGTTTCCTCATCCATATAGTAAGCCTAGTGAATGCTTTTT  2760
  Cys  Met  Leu  Ile  Ser  Met  Gln  Phe  Pro  His  Pro  Tyr  Ser  Lys  Pro  Ser  Glu  Cys  Phe  Leu   139
ACTATATCCATTTAGATTAATTTTTAAACTGCTTTTAGATAAACGCCTACAAGGTAGGCT  2820
  Leu  Tyr  Pro  Phe  Arg  Leu  Ile  Phe  Lys  Leu  Leu  Leu  Asp  Lys  Arg  Leu  Gln  Gly  Arg  Leu   159
ATATCATTATGAAGTATATAAAATTATTATTCATACAATATCTATAGATGAAGCCAAATA  2880
  Tyr  His  Tyr  Glu  Val  Tyr  Lys  Ile  Ile  Ile  His  Thr  Ile  Ser  Ile  Asp  Glu  Ala  Lys  Tyr   179
TGAATTTCTAGTCAAGAGTATATTAAATTCTAGAAAAAAATCTTGGAATGAGAAACTTAA  2940
  Glu  Phe  Leu  Val  Lys  Ser  Ile  Leu  Asn  Ser  Arg  Lys  Lys  Ser  Trp  Asn  Glu  Lys  Leu  Asn   199
TGAATTATCAGAAATACAACATAAAGTTGTTAAATCTGTATATGAGTGGCAATACTACAT  3000
  Glu  Leu  Ser  Glu  Ile  Gln  His  Lys  Val  Val  Lys  Ser  Val  Tyr  Glu  Trp  Gln  Tyr  Tyr  Ile   219
TGTGCCATTGTTAGGTAGCTTACACATTTTTAAAATTAATAATGGAGATATAGAGCAAAA  3060
  Val  Pro  Leu  Leu  Gly  Ser  Leu  His  Ile  Phe  Lys  Ile  Asn  Asn  Gly  Asp  Ile  Glu  Gln  Lys   239
ACTTTATCATCCTCAAAAAGATGGAAGTAAATCACCTCCAACAGCAAGAAAGGCAAATAA  3120
  Leu  Tyr  His  Pro  Gln  Lys  Asp  Gly  Ser  Lys  Ser  Pro  Pro  Thr  Ala  Arg  Lys  Ala  Asn  Asn   259
TGGCTATGTAGAAATTAATGATAATTTGACTAATTTTATTGATAAATTATTGAACAAGTA  3180
  Gly  Tyr  Val  Glu  Ile  Asn  Asp  Asn  Leu  Thr  Asn  Phe  Ile  Asp  Lys  Leu  Leu  Asn  Lys  Tyr   279
TTCTTTTTTAGATACCCCTATCTTATTATCTGATAGTCAACGAAAATCAAATGATGTGAC  3240
  Ser  Phe  Leu  Asp  Thr  Pro  Ile  Leu  Leu  Ser  Asp  Ser  Gln  Arg  Lys  Ser  Asn  Asp  Val  Thr   299
TAAGGAAATTTATTCTTTTTACCCAGAGCTTCTTTTAGCAGAAATTGGCGAAACAATCTC  3300
  Lys  Glu  Ile  Tyr  Ser  Phe  Tyr  Pro  Glu  Leu  Leu  Leu  Ala  Glu  Ile  Gly  Glu  Thr  Ile  Ser   319
TTTTGAATCTCATATTTTAAATATTCCTAAATTAATCACAGAATATTCCAAGAATCCAGA  3360
  Phe  Glu  Ser  His  Ile  Leu  Asn  Ile  Pro  Lys  Leu  Ile  Thr  Glu  Tyr  Ser  Lys  Asn  Pro  Asp   339
TAATAGCACATCTGGCAAATTTGAGAAAATTCTTGAGGAAGCTTTTAATTTATTCATTGA  3420
  Asn  Ser  Thr  Ser  Gly  Lys  Phe  Glu  Lys  Ile  Leu  Glu  Glu  Ala  Phe  Asn  Leu  Phe  Ile  Asp   359
TGTAGAAGCTCAATGGTTAGCAGGAGCAGGCAGAACAGATATTGAATGTATGTATTTGCC  3480
  Val  Glu  Ala  Gln  Trp  Leu  Ala  Gly  Ala  Gly  Arg  Thr  Asp  Ile  Glu  Cys  Met  Tyr  Leu  Pro   379
AATAAATGAAAAATTTTCTATTGAAGCTAAATCAACAAAAAATAAATTAAGTATGATTAA  3540
  Ile  Asn  Glu  Lys  Phe  Ser  Ile  Glu  Ala  Lys  Ser  Thr  Lys  Asn  Lys  Leu  Ser  Met  Ile  Asn   399
TTCAGGTAGATTAAAACGCCACAGAACTTTAATTAGTGCAAATTATACTATTGTTATTAC  3600
  Ser  Gly  Arg  Leu  Lys  Arg  His  Arg  Thr  Leu  Ile  Ser  Ala  Asn  Tyr  Thr  Ile  Val  Ile  Thr   419
TCCAAGATATGTACCTAGTGTACGTTATGATATTGAAGCACAAGATATAGTTCTCATAAC  3660
  Pro  Arg  Tyr  Val  Pro  Ser  Val  Arg  Tyr  Asp  Ile  Glu  Ala  Gln  Asp  Ile  Val  Leu  Ile  Thr   439
AGCCGATACATTAGCAGAGTATTTATACAATAATATTATTTCTAATAATCGAGATATTTC  3720
  Ala  Asp  Thr  Leu  Ala  Glu  Tyr  Leu  Tyr  Asn  Asn  Ile   Ile  Ser  Asn  Asn  Arg  Asp  Ile  Ser   459
TTATGCTGATATACAAGCTATCATTGTTGCTAATTTGGGAAAAGATATTAGTACTCAAAT  3780
  Tyr  Ala  Asp  Ile  Gln  Ala  Ile  Ile  Val  Ala  Asn  Leu  Gly  Lys  Asp  Ile  Ser  Thr  Gln  Ile   479
TTCAAATTTAACACTTTCCAAATTTGGTTGATGGATAAATGAATATGCATAAGAACATTA  3840
  Ser  Asn  Leu  Thr  Leu  Ser  Lys  Phe  Gly                                                       488
CAGAAAAAACTATTGTATGATAGCAATCACAAAATTGTGATCTACATACCTCATAACTTA  3900
TACAGTTTTAAAAGGGTTAAATCGCCATTCCTATATTCGAGTGGGCGATTTTTTATGTCT  3960
AGCGTTTAGCTTTGAGTACCTTGTACACACAAGGAATACTTATGGATCATTACTTACTTT  4020
CAAGCAAATCTAAAACATTATCCCTTAAACATATCTTCCGTCTAAGTGAAGATGAAGCCT  4080
TCCAACTTTTGAAAGCAAACCGTTGGGAAAATCCTGATAAAATCCTCCAAAGTAGTACGT  4140
CTAAATTATAGAATAAGCCACTTAAGATAATCACTCAAGGTGTTTAAAATATCATCAATT  4200
AAACTTGTCAAAAAATTTCTCTTGACAGAACTAAGACCTTTTTGGCACTCGCTATATTCA  4260
AAAATGCAGAAAATTTCAAAAAGCACTTGACAAAAGAAAACGCTTGACAAGCACCCTATA  4320
AAATAATATCTTTTTGCCCACCGAGAATAGTCCCTTCATTATGACTTTGTTTTTTAAGCC  4380
TAAAAAGGCTTGGAAATTTCAGAGTTTTAATGAAGGGATTTTTGTTTCTATGGCACAACA  4440
TTTTCGGTTATCAGCGACAGCACGCCAGCTATCTATGGATAGCTTGGTGAGCTTGTCTGA  4500
TGACGAGATTTTTCAGCTATTGAAACAAGCTCGATGGGGTAATGTAGATGAAATAAACGA  4560
TGTTATTTGTCCGCATTGCCATACTCGCCACCACGCCTATTTCATTTCTACTCGTAAACA  4620
ATGGCAGTGCAAACATTGCCAACATCGTTTTTCTATTACAGCAGGCACTATTTTTCAAGG  4680
TGCAAAACTTTCTTTGCGTAAAATTGTATTGGCAATTTTCTATTTTTCGACTGAAAGCAA  4740
AGGATTATCTGCTATTACACTATCACACAAGCTCAACGTGCAATATAAAACAGCGTGGGT  4800
ATTGCTACATAAATTTCGTGAATCGCTTGATAAAGCCAAAGATTTAACACATTTAAGTGG  4860
CGAAGTTCATATTGATGGCGCTTACATGAACAACTACATTCGTCCTAAAAACTTTCTCCA  4920
TAAACGTATTGATAGATGAAAGAAACGTTATCAGCGTGCAGATAAATCTTGTGTATTGGT  4980
ATTCCGTC                                                       4988
```

Fig. 1 Nucleotide sequence of the leftmost *Cla*I to *Hin*cII region of the 4.5 kb insert in pKS456. The amino acid sequence assigned for R*Hga*I are given below the nucleotide sequence. The nucleotide sequence is numbered from the leftmost end of pKS318[9], and the amino acid sequences from the first initiation codon, respectively.

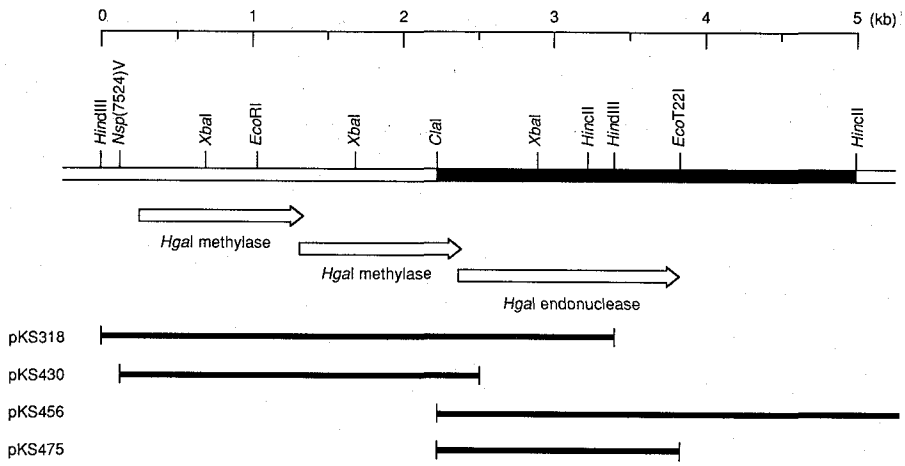

Fig.2. Gene organization of the *Hga*I R-M system. The *filled line* represents the region of which the DNA sequence was determined as shown in Fig.1 and *open line* the other region for *H.gallinarum* chromosomal DNA. Below the physical map on which the major restriction sites are indicated, the two methylase and one endonuclease genes are shown by *open arrows*, and the region inserted into the plasmid costruct in this experiment are indicated by *bold lines*.

in the appropriate position.To confirm this assignment, I have a plan to analyze the N-terminal amino acid sequence.

The gene organization of the *Hga*I R-M system derived from Ref.10 and this paper is shown in Fig.2. This R-M system is consisted of two cytosine methylase genes responsible for modification of different DNA strands in the target DNA and a endonuclease gene for cleavage of both DNA strands.

Acknowlegments

I wish to thank Dr. M. Takanami for his instructive advice and criticism during the course of this work.

References

(1) G.G. Wilson, *Nucleic Acids Res.*, **19**, 2539-2566(1991).
(2) G.G. Wilson, and N.E. Murray, *Ann. Rev. Genetics*, **25**, 585-627(1991).
(3) T.A. Bickle, and D.H. Krüger, *Ann. Rev. Microbiol.*, **57**, 434-450(1993).
(4) K. Kita, H. Kotani, H. Sugisaki, and M. Takanami, *J. Biol. Chem.*, **264**, 5751-5756(1989).
(5) H. Bocklage, K. Heeger, and B. Müller-Hill, *Nucleic Acids Res.*, **19**, 1007-1013(1991).
(6) K. Kita, M. Suisha, H. Kotani, H. Yanane, and N. Kato, *Nucleic Acids Res.*, **20**, 4167-4172 (1992).
(7) K. Kita, H. Kotani, H. Ohta, H. Yanase, and N. Kato, *Nucleic Acids Res.*, **20**, 618(1992).
(8) N.L. Brown, and M. Smith, *Proc. Natl. Acad. Sci. USA*, **74**, 3213-3216(1977).

(9) H. Sugisaki, *Gene (Amst.)*, **3**, 17-28(1978).
(10) H. Sugisaki, K. Yamamoto, and M. Takanami, *J. Biol. Chem.*, **266**, 13952-13957(1991).
(11) C. Yanisch-Perron, J. Vieira, and J. Messing. *Gene (Amst.)*, **33**, 103-119(1985).
(12) A.C.Y. Chang, and S.N. Cohen, *J. Bacteriol.*, **134**, 1141-1156(1987).
(13) J. Sambrook, E.F. Fritsch, and T. Maniatis, "Molecular cloning: A Laboratotry Manual" 2nd Ed, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
(14) F. Sanger, S. Nicklen, and A.R. Coulson, *Proc. Natl. Acad. Sci. USA*, **74**, 5463-5467(1977)
(15) J. Messing, *Methods in Enzymol.*, **101**, 20-78(1983).