Frequency and Distribution of DNA Uptake Signal Sequences in the *Haemophilus influenzae* Rd Genome

Hamilton O. Smith,* Jean-Francois Tomb, Brian A. Dougherty, Robert D. Fleischmann, J. Craig Venter

The naturally transformable, Gram-negative bacterium *Haemophilus influenzae* Rd preferentially takes up DNA of its own species by recognizing a 9–base pair sequence, 5′-AAGTGCGGT, carried in multiple copies in its chromosome. With the availability of the complete genome sequence, 1465 copies of the 9–base pair uptake site have been identified. Alignment of these sites unexpectedly reveals an extended consensus region of 29 base pairs containing the core 9–base pair region and two downstream 6–base pair A/T-rich regions, each spaced about one helix turn apart. Seventeen percent of the sites are in inverted repeat pairs, many of which are located downstream to gene termini and are capable of forming stem-loop structures in messenger RNA that might function as signals for transcription termination.

Haemophilus influenzae Rd is a Gram-negative bacterium whose complete genome sequence is now known (*1*). It is a naturally transformable organism that preferentially takes up double-stranded DNA of its own species from the medium and integrates one strand into its chromosome to yield genetic transformants. Recognition and efficient uptake of homospecific donor DNA molecules is facilitated by the presence of uptake signal sequences (USSs) in the donor molecules (*2*–*5*). It is now possible, using the genome sequence (*1*), to examine the frequency and distribution of these uptake sites with a precision that was heretofore impossible.

The USS sites were originally identified as an 11–base pair (bp) sequence, 5′-AAGTGCGGTCA, common to four small *H. parainfluenzae* donor DNA fragments showing preferential uptake from a mixture of restriction fragments (*4*, *5*). Subsequent examination of additional USSs in *H. influenzae* DNA (*6*) and *Haemophilus* phage DNA (*7*) showed that only the first nine residues were highly conserved. Uptake competition experiments (*3*) and Southern (DNA) blot analyses of *H. influenzae* DNA with an oligonucleotide USS probe (*8*) revealed the presence of several hundred USSs in the genome. Analysis of a contiguous 9.1-kb region of the *H. influenzae* genome (*9*) revealed 14 USSs, eight in the plus orientation (5′-AAGTGCGGT) and six in the minus orientation (5′-ACCGCACTT). Two pairs of USSs formed inverted-repeat, stem-loop configurations just downstream of the 3′-terminus of genes, and eight of the other sites were in coding regions. The naturally transformable Gram-negative bacterium *Neisseria gonorrhoeae* also has DNA USSs. These 10-bp sites have the sequence 5′-GCCGTCTGAA and occur frequently in paired stem-loop configurations immediately downstream to gene termini (*10*).

A total of 1465 copies of the USS were found by searching the complete genome sequence of *H. influenzae* Rd (Fig. 1). Taking into account the 62% A/T base composition of *Haemophilus*, only about eight USSs would be expected to occur by chance (*11*). The USSs were distributed largely at random over the genome with 734 in the plus orientation and 731 in the minus orientation. However, the distribution was not entirely random because only 65% of the sites were found in open reading frames, whereas about 86% of the genome is coding sequence, excluding transfer RNA and ribosomal RNA genes (*1*). There was no obvious polarity of plus- or minus-oriented sites in any region of the genome; the longest tracts of sites in any particular orientation were only seven to eight in length, and the distribution of tract lengths for both plus and minus sites appeared to be random.

Figure 2A shows the percentage occurrence of each of the four bases at positions flanking the core 9-bp sequence, when all 1465 copies were aligned in the plus direction. A 29-bp consensus USS was identified that has the sequence 5′-aAAGTGCGGT.rwwwww. . . . . rwwwww, in which uppercase letters represent conserved bases, lowercase letters are bases that occur in >50% of the USSs, a dot is any base, r is purine, and w is A or T. Previous work has shown that the introduction of bulky ethyl groups onto individual phosphate groups in the DNA backbone at certain positions of the site interferes with recognition of the DNA during transformation, as indicated by asterisks in Fig. 2A (*4*). Interference occurs well beyond the borders of the 9-bp core site, a result that originally was difficult to explain but now makes sense on the basis of the extended site. The original region of analysis did not include the second rwwwww repeat because it was unanticipated that the USS would be so large. The large number of USSs in the genome has raised the question whether mutational drift of some of the sites may have occurred. Figure 2B shows the base frequencies for 764 USSs mutated at single positions in the conserved 9-bp core region. The extended site region is still visible, although the background contributed by the approximately 254 singly mutated sites expected by chance (*12*) blurs the consensus pattern consider-

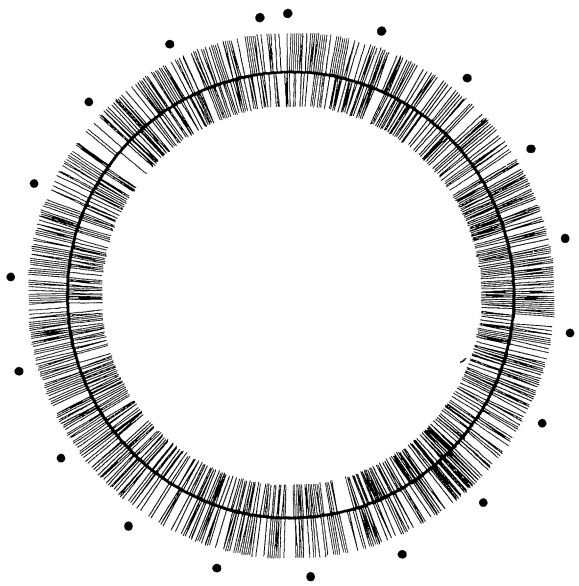

Fig. 1. Location of DNA uptake signal sequences on the *H. influenzae* Rd genome. Dots indicate scale in 100-kb units. The single Not I site (*18*) is at zero (indicated by dot at 12 o'clock), and the sequence is oriented as in (*1*). The 734 plus sites (5′-AAGTGCGGT) are outside, and the 731 minus sites (5′-ACCGCACTT) are inside.

H. O. Smith, J.-F. Tomb, B. A. Dougherty, Department of Molecular Biology and Genetics, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA.
R. D. Fleischmann and J. C. Venter, The Institute for Genomic Research, 932 Clopper Road, Gaithersburg, MD 20878, USA.

*To whom correspondence should be addressed.

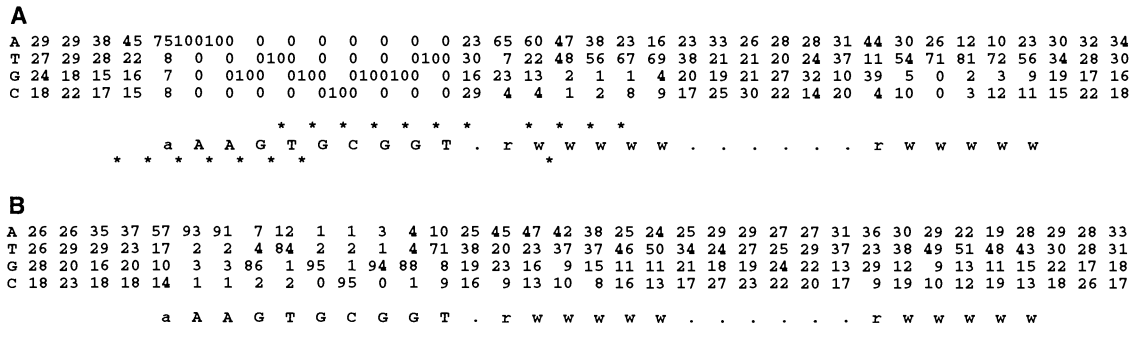

Fig. 2. Alignment of USSs in the plus orientation. Numbers (rounded to lower whole number) are percent base frequencies at each position in the order (top to bottom) A, T, G, and C. The consensus sequence is given below each block. (**A**) Alignment of 1465 sites with complete conservation of the core 9-bp sequence. Asterisks above (5′-3′ strand) and below (3′-5′ strand) the consensus sequence indicate phosphate groups that, when individually ethylated, interfere with DNA uptake (*4*). (**B**) Alignment of 764 sites with a single base change in the core sequence.

ably. If the lengths of all the 29-bp USSs and the singly mutated sites are added, there is a total of 2229 sites occupying approximately 3.5% of the genome.

The mean distance between sites was 1248 bp. The distribution of intersite distances on the genome was essentially random over the range of 50 bp to 8 kb. Only four intersite distances were greater than 8 kb. Two regions of 14,441 and 8293 bp that contained no sites were located in the interval from 1.56 to 1.59 Mb on the genome (Fig. 1). This is in a region of relatively high G/C content showing similarity to a mu-like phage (*1*). It is not surprising that this foreign DNA contained few USS sites. A more striking deviation from randomness occurred at very close intersite distances. There were 127 pairs of sites with a separation of ≤35 bp between the starts of each site, representing 17.3% of the total sites. By chance alone, one would expect only 39 such pairs. The spacings were nonrandom, occurring in two major peaks at separations of 17 ± 2 bp (64 cases) and 30 ± 2 bp (38 cases). Instead of a roughly equal number of these pairs in each possible order (plus/plus, plus/minus, minus/plus, or minus/minus), the distribution was biased, with 87 in a plus/minus stem-loop configuration, 39 in a minus/plus stem-loop configuration, 1 in a plus/plus pair configuration, and 0 in a minus/minus configuration.

Eighty-two percent of the minus/plus stem-loop pairs and 61% of the plus/minus stem-loop pairs were wholly in intergenic regions. The stem-loop pairs were frequently located downstream to the termination of a gene (Fig. 3). By analogy with *Escherichia coli*, they could act as signals for transcription termination. Those inverted repeat pairs contained within coding regions or overlapping the starts or ends of coding regions could possibly have a regulatory function or influence the stability of the messenger RNA.

Of the plus/minus pairs, 46 had spacers exactly 8 bp in length between the two USSs (Fig. 3). This separation preserves most of the 29-bp consensus for each USS, so that each member of the pair can be an active site. This is illustrated in Fig. 4A, which shows that the rwwwww regions coincide. Overlap with the G/C-rich portion of the core sequence cannot occur because of sequence incompatibility. Thirty-four of the plus/minus pairs had spacers of 19 to 22 bp (Fig. 3), which again favors the retention of the 29-bp consensus for each of the sites in the pair (Fig. 4A). The preferred 8-bp separation may be related to the greater stability of stem-loop structures with a smaller loop size. The separation distances for the minus/plus pairs were more variable than for the plus/minus pairs, which is as expected because there is no problem with sequence overlap incompatibility when the rwwwww repeats are directed outward (Fig. 4B).

The frequent occurrence of USSs as inverted repeats in intergenic regions is reminiscent of the 500 to 1000 REP (repetitive extragenic palindrome) sequences found in *E. coli* and *Salmonella typhimurium* DNA (*13*). These conserved, palindromic sequences can form stem-loop structures in mRNA and appear either alone or in up to four tandem inverted repeat copies within intergenic regions of many operons. They do not act as terminators or ribonuclease III processing sites but have been shown to stabilize upstream mRNA by forming ener-

Fig. 3. Some examples of USS pairs downstream to the termini of genes that have the potential to form stem-loops in mRNA. (**A**) Plus/minus stem-loop pairs with 19- to 22-bp spacings. (**B**) Plus/minus stem-loop pairs with 6- to 8-bp spacings. (**C**) Minus/plus stem-loop pairs with variable spacings. Gene names are followed by the termination triplet (uppercase) for that gene and 3′ downstream sequence containing the USS inverted repeat pair (uppercase). For comparison purposes, the first USS of each pair is aligned by deleting several bases (-x-) or by adding spaces. Complementary regions capable of forming the stem of the stem-loop structure in mRNA are underlined.

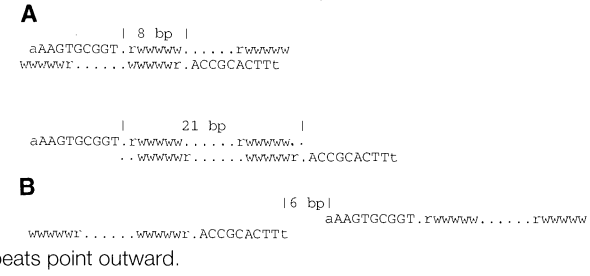

Fig. 4. Allowable USS pair spacings that permit retention of the 29-bp consensus sequence. (**A**) Plus/minus pairs in 8-bp spacing and 21-bp spacing. By mentally sliding the two sites with respect to each other, one can see which spacings avoid interference between the conserved regions of the sites. (**B**) Minus/plus pairs. Any spacing is permissible because the rwwwww repeats point outward.

getically stable stem-loop structures resistant to processive 3′ exonucleases (*13*). Biochemical analysis of USS stem-loops is needed to determine their potential role in mRNA stabilization or termination.

USSs are in genetic equilibrium with mutated sites containing one or more mismatches. In the absence of selection, the frequencies of mutated and nonmutated sites are simply those found in a random sequence. Any observed excess of sites must be accounted for by some form of selection. In *H. influenzae*, selection for donor DNA containing USS occurs at the cell surface. Thus, one restoring force for correct sites is transformation itself (*14*). If a cell in the population loses a site to mutation, that site will tend to be replaced with a correct site by transformation, because donor DNA carrying the correct site is preferentially taken up compared to donor DNA carrying the incorrect version of the site. An additional selective advantage might derive from the participation of a significant fraction of the sites in stem-loop structures with possible roles in transcription termination or regulation. A third selective advantage might come from any role that the USSs might play as recombinational hotspots, similar to the χ sites of *E. coli* (*15*).

The χ sites have the sequence 5′-GCTGGTGG (plus orientation) in *E. coli* and are recognized by *recBCD* exonuclease V (*15*). Exonuclease V moves processively along the DNA, unwinding and cleaving the DNA until a χ site is encountered in the minus orientation. The enzyme then cleaves near the χ site and undergoes a structural change such that further cleavage is suppressed and the strands are unwound, producing a free 3′ single strand that can synapse with homologous DNA to initiate recombination with the help of the RecA protein (*15*). In *E. coli*, the sites are distributed with a strong strand bias such that moving clockwise from the origin of replication, the sites are mostly in the plus orientation and counterclockwise they are mostly in the minus orientation (*15*). The average spacing of χ sites in *E. coli* is 5 kb (*15*). *H. influenzae* has genes homologous to the *recB*, *recC*, and *recD* genes of *E. coli* (*1*), and the *H. influenzae* exonuclease V has been purified and extensively studied (*16, 17*). Its properties are similar to those of the *E. coli* enzyme. USSs are frequent but lack the regional strand bias characteristic of the *E. coli* χ sites. The plus and minus sites appear to be randomly mixed (Fig. 1). Runs of plus USS sites or of minus USS sites do not exceed eight repeats in length, and the distribution of run lengths is about as expected by chance. On the other hand, a search of the *H. influenzae* genome reveals 98 copies of the sequence 5′-GCTGGTGG, 44 in the plus orientation and 54 in the minus orientation, and only eight would be expected in each orientation by chance. However, only a weak strand bias of these putative plus and minus χ sites is seen relative to the origin. There are eight plus putative χ sites and 22 minus putative χ sites in 600 kb to the left of the *ori* (origin of replication) site, located at position 602 kb on the genome sequence (*1*), and 18 plus versus 21 minus putative χ sites in 600 kb to the right of *ori*. Whether *H. influenzae* and *E. coli* share the same χ site specificity will have to be determined by complementation of *recBCD* mutants of *E. coli* with *H. influenzae* genes.

REFERENCES AND NOTES

1. R. Fleischmann *et al.*, *Science* **269**, 496 (1995).
2. J. J. Scocca, R. L. Poland, K. C. Zoon, *J. Bacteriol.* **118**, 369 (1974).
3. K. L. Sisco and H. O. Smith, *Proc. Natl. Acad. Sci. U.S.A.* **76**, 972 (1979).
4. D. B. Danner, R. A. Deich, K. L. Sisco, H. O. Smith, *Gene* **11**, 311 (1980).
5. D. B. Danner, H. O. Smith, S. A. Narang, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 2393 (1982).
6. S. H. Goodgal and M. A. Mitchell, *J. Bacteriol.* **172**, 5924 (1990).
7. W. P. Fitzmaurice, R. C. Benjamin, P. C. Huang, J. J. Scocca, *Gene* **31**, 187 (1984).
8. M. E. Kahn and H. O. Smith, *J. Membr. Biol.* **81**, 89 (1984).
9. J.-F. Tomb, H. el-Hajj, H. O. Smith, *Gene* **104**, 1 (1991).
10. S. D. Goodman and J. J. Scocca, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 6982 (1988).
11. The nucleotide frequency of *H. influenzae* Rd is A = 0.308, T = 0.310, G = 0.192, and C = 0.190. The expected random frequency of the site AAGTGCGGT per genome is approximately $(0.31)^4 \times (0.19)^5 \times 1830121 = 4.2$. The expected frequency for both orientations of the site is then 8.4 per genome.
12. The expected random occurrence of a singly mutated USS is

$$[8 \times (0.19)^6(0.31)^3 + 9 \times (0.19)^5(0.31)^4 + 10 \times (0.19)^4(0.31)^5] \times 1{,}830{,}121 \times 2 = 254$$

13. C. F. Higgins, R. S. McLaren, S. F. Newbury, *Gene* **72**, 3 (1988).
14. R. J. Redfield, *Nature* **352**, 25 (1991).
15. G. R. Smith, *Experientia* **50**, 234 (1994); V. Burland, G. Plunkett III, D. L. Daniels, F. R. Blattner, *Genomics* **16**, 551 (1993).
16. E. A. Friedman and H. O. Smith, *J. Biol. Chem.* **247**, 2846 (1972).
17. K. W. Wilcox and H. O. Smith, *ibid.* **251**, 6127 (1976).
18. L. Kauc, M. A. Mitchell, S. H. Goodgal, *Gene* **95**, 149 (1990).

19. H.O.S. is an American Cancer Society Research Professor. This work was supported in part by research grants from the American Cancer Society and NIH. B.A.D. is supported by an NIH Training Grant.

16 May 1995; accepted 28 June 1995

Effects of the *obese* Gene Product on Body Weight Regulation in *ob/ob* Mice

Mary Ann Pelleymounter,* Mary Jane Cullen, Mary Beth Baker, Randy Hecht, Dwight Winters, Thomas Boone, Frank Collins

C57BL/6J mice with a mutation in the *obese* (*ob*) gene are obese, diabetic, and exhibit reduced activity, metabolism, and body temperature. Daily intraperitoneal injection of these mice with recombinant OB protein lowered their body weight, percent body fat, food intake, and serum concentrations of glucose and insulin. In addition, metabolic rate, body temperature, and activity levels were increased by this treatment. None of these parameters was altered beyond the level observed in lean controls, suggesting that the OB protein normalized the metabolic status of the *ob/ob* mice. Lean animals injected with OB protein maintained a smaller weight loss throughout the 28-day study and showed no changes in any of the metabolic parameters. These data suggest that the OB protein regulates body weight and fat deposition through effects on metabolism and appetite.

Mutation of the *obese* gene in the C57BL/6J mouse results in a syndrome that includes obesity, increased body fat deposition, hyperglycemia, hyperinsulinemia, and hypothermia (*1*). Parabiosis studies have suggested that the mutant obese mouse (*ob/ob*) lacks a blood-borne factor that could regulate adiposity by modulation of appetite and metabolism (*2*). Here we test the hypothesis that the recently cloned *obese* gene (*3*) is involved in the regulation of adiposity by administering the OB protein to *ob/ob* mice.

The OB protein was expressed in *Escherichia coli* and purified to homogeneity as a 16-kilodalton monomer (*4*). The protein was dissolved in phosphate-buffered saline (PBS) (pH 7.4) and administered by daily intraperitoneal injection (0.1, 1.0, or 10.0 mg/kg) to 5-week-old C57BL/6J mice that were either homozygous (*ob/ob*) or heterozygous (+/?) for the *obese* gene mutation. The OB protein was also administered to 8-week-old, weight-stabilized normal C57BL/6J mice. Controls received equivolume (10 ml/kg) injections of PBS. Body weight, food

M. A. Pelleymounter, M. J. Cullen, M. B. Baker, F. Collins, Department of Neurobiology, Amgen, Inc., 1840 De-Havilland Drive, Thousand Oaks, CA 91320, USA.
R. Hecht, D. Winters, T. Boone, Department of Recovery Process Development, Amgen, Inc., Thousand Oaks, CA 91320, USA.

*To whom correspondence should be addressed.