Organization of the *Escherichia coli* and *Salmonella typhimurium* chromosomes between flagellar regions IIIa and IIIb, including a large non-coding region

MANIDIPA RAHA, MAY KIHARA, IKURO KAWAGISHI† and ROBERT M. MACNAB*

Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06511, USA

(Received 15 March 1993; revised 15 April 1993; accepted 26 April 1993)

Flagellar regions IIIa and IIIb of the *Escherichia coli* and *Salmonella typhimurium* chromosomes (at 40 min and 42-43 min, respectively) has been shown to be separated by DNA unrelated to flagellar function, with region IIIa being immediately followed by a gene, *amyA*, that encodes a cytoplasmic α-amylase. The chromosome between *amyA* and flagellar region IIIb has now been investigated. The high level of DNA similarity between the *E. coli* and *S. typhimurium* sequences that exists in flagellar region IIIa and in *amyA* continues initially, with three genes of unknown function; in *E. coli*, there may be a fourth gene. The remainder of the region, up to the start of flagellar region IIIb, lacks any obvious open reading frames, scores poorly on an algorithm for coding probability, has a high A+T content, and is totally dissimilar in the two species. We conclude that it is non-coding. In *E. coli* this region extends for 2-7 kb and in *S. typhimurium* for 0-8 kb. These values are unusually large for prokaryotes, where the non-coding regions between operons are generally quite short. The data, which are discussed in the context of a hypothesized disruption of a contiguous ancestral flagellar region, may give new insight into the organization and evolution of the bacterial chromosome.

Introduction

In contrast to eukaryotic genomes, bacterial genomes are compact, with typically only around 50-300 bp of non-coding sequence between adjacent operons. For example, in the flagellar regulons of *Salmonella typhimurium* and *Escherichia coli*, (subscripts E and S are used throughout this paper to distinguish *E. coli* and *S. typhimurium* genes) among the cases where the non-coding distance between adjacent operons has been established, the mean value is 151 bp, the highest value is 263 bp (between the *fliC* and *fliD* operons of *E. coli*; Hanafusa et al., 1989), and the lowest value is 64 bp (between the *flgB* and *flgK* operons of *S. typhimurium*; Homma et al., 1990).

We showed recently (Kawagishi et al., 1992) that a set of 19 flagellar genes in *E. coli*, which originally had been called flagellar region III, in fact consists of two contiguous sets (region IIIa, containing 5 genes, and region IIIb, containing 14 genes) separated by about 6-6 kb; a similar situation exists in *S. typhimurium* except that the intervening region is smaller (4-4 kb). Thus the last operon in flagellar region IIIa and the first operon in flagellar region IIIb are separated by a large distance. What is the nature and function of this DNA? Chromosomal deletion established that it contains no genes necessary for motility or for normal growth rates on minimal medium, but of course this does not exclude the possibility that it contains non-essential genes. Indeed, this intervening region is so large that it would be surprising if it was entirely non-coding.

The supposition that it would contain at least some coding sequence was soon vindicated by the finding that the last flagellar operon in region IIIa (the *fliD* operon) was immediately followed by a gene, *amyA*, which encodes a cytoplasmic α-amylase (Raha et al., 1992). We have now examined the entire intervening region, and conclude that although it does contain several genes, it also contains an unusually long stretch of non-coding sequence.

Methods

Bacterial strains. *E. coli* strains DH5α (Gibco BRL) and XL1-Blue and JM101 (Stratagene) were used for cloning, subcloning and M13 production.

*Author for correspondence. Tel. (203) 432 5590; fax (203) 432 5592; e-mail rmacnab@yalevm.yec.yale.edu.

† Present address: Department of Molecular Biology, Nagoya University, Nagoya, Japan.

The nucleotide sequence data reported in this paper have been submitted to GenBank and have been assigned the accession numbers L13279 and L13280.
Cloning and sequencing of $E$. coli DNA. A plasmid with an insert that contains the fliDE operon of flagellar region IIIa, the entire intervening region, and flagellar region IIIb genes from fliE to fliF through an EcoRI site in fliG (Kawagishi et al., 1992). A 2.1 kb fragment of the insert between the MluI and XbaI sites was subcloned into pIB25 (International Biotechnologies Inc.) to give plasmid pMR7. Using the XbaI site and the vector HindIII site, the cloned fragment from pMR7 was recloned into the corresponding sites in plasmid Bluescript KS (Stratagene) to give plasmid pMR8.

The approaches used to sequence amyA, and the DNA up to a MluI site 2.2 kb beyond amyA have been described (Raha et al., 1992), as have those for the DNA from an XbaI site 0.7 kb before fliE through that gene (Müller et al., 1992). Exonuclease III deletions were generated from both ends of the inserts of pMR7 and pMR8, using the XbaI and KpnI sites of pMR7 and the HindIII and KpnI sites of pMR8; these were then used for sequencing by the dideoxy chain termination method using commercial synthetic primers. Both the MluI and the XbaI sites were sequenced through, using appropriate plasmids.

Cloning and sequencing of $S$. typhimurium DNA. oYAl is a pUC119-based plasmid (Müller et al., 1992); it has a 6.7 kb EcoRI-EcoRI insert of $S$. typhimurium DNA that extends from within fliD in flagellar region IIIa through all of the intervening region and as far as fliF in flagellar region IIIb.

The approaches used to sequence amyA, and fliE have been described (Raha et al., 1992; Müller et al., 1992). For sequencing the rest of the intervening region, four M13 derivatives of oYAl1 were used. The first two contained a ClaI-HpaI fragment (from within amyA, to near the end of orf48), cloned into M13mp18 and M13mp19, while the second two contained a BamHI-EcoRI fragment (from within orf48 to within fliF) cloned into the same vectors.

Chemicals and enzymes. All chemicals and enzymes were obtained from standard commercial sources. Antibiotics were used at the following concentrations: ampicillin, 50 µg ml$^{-1}$; tetracycline, 10 µg ml$^{-1}$. Synthetic primers (18-mers) were synthesized using an Applied Biosystems 392 DNA/RNA Synthesizer.

Results

Overall organization of the intervening regions in E. coli and S. typhimurium

The fliD site in flagellar region IIIa is followed by an amylase gene, amyA, oriented clockwise on the chromosome (Raha et al., 1992). We have now extended the sequence analysis all the way through the intervening region to the 3' end of fliE (see Methods for details). The results are shown schematically in Fig. 1(a); the corresponding DNA sequence is available in GenBank under accession number L13287.

The fragment used for analysing amyA contained a second open reading frame oriented counterclockwise, with a short (77 bp) distance between the 3' ends of the respective coding regions. This second open reading frame has a deduced product molecular mass of 15 kDa, and will be referred to as orf15. Diverging from the 3' non-coding region of orf15 were two open reading frames with an ATGA overlap between them; both had strong consensus ribosome binding sequences. They correspond to products with deduced molecular masses of 48 and 9 kDa, respectively, and so will be referred to as orf48 and orf9. About 100 bp following the 3' end of orf9 was a fourth short open reading frame, orf8, with an extremely strong, well-placed consensus ribosome binding sequence (AGGAGG), and followed by a pronounced stem-loop sequence. Based on their opposite orientation, orf48 and orf9 presumably belong to distinct transcriptional units. With a 4 bp overlap of their coding regions, orf48 and orf9 must belong to the same transcriptional unit, while the distance between the 3' end of orf9 and the 5' end of orf8 suggests that the latter is a separate transcriptional unit. It is not obvious from inspection what the corresponding promoters are and, in the absence of knowledge regarding gene product function (see below), we did not attempt to determine the transcription start sites.

For the next 2-7 kb, until the first flagellar gene in region IIIb, fliE, there was no evidence for any further genes: open reading frames in both orientations were very short (mostly less than 300 bp) and had no obvious ribosome binding site.

The intervening region of $S$. typhimurium followed a similar pattern (Fig. 1(b); sequence available in GenBank under accession number L13280), with two exceptions: there was no evidence for an open reading frame corresponding to orf8, and the distance between the last open reading frame, orf9, and fliE was much shorter (0-8 kb) than in $E$. coli.

Comparison of the DNA sequence of the intervening region in the two species

When DNA sequences of the intervening region in $E$. coli and $S$. typhimurium were compared in a matrix plot (Fig. 2), the similarity seen previously in flagellar region IIIa and amyA was found to extend through orf15, orf48 and orf9. These open reading frames in the two species show 77, 81 and 86% identity, respectively, when aligned by the algorithm of Lipman & Pearson (1985).

Following orf9, no sequence similarity was discerned, either by inspection of Fig. 2 or by attempted alignment, until the start of flagellar region IIIb. Thus orf8 plus the 2-7 kb between orf8 and fliE is unrelated to the 0-8 kb between orf9 and fliE. If they were part of the same sequence originally, they must since have diverged beyond recognition, and also (at least in the case of $S$. typhimurium) undergone further rearrangement by deletion.

Evidence that orf15, orf48 and orf9 are authentic genes

Minicell analysis (Raha et al., 1992) has demonstrated that orf15 encodes a 15 kDa protein and so is an authentic gene. We conclude that orf48 and orf9 are
authentic genes also, based on the following considerations.

The strongest evidence is that the deduced amino acid sequences show even higher levels of identity than the corresponding nucleotide sequences (83, 92 and 100% for Orf15, Orf48 and Orf9, respectively; Fig. 3). This result is statistically extremely improbable for non-coding DNA, because of the much larger basis set of 20 amino acids vs. that of four nucleotides. Inspection of the genetic code reveals that a one-base codon change has a probability of 0.76 of causing an amino acid change, and so between closely related sequences (where changes of more than one base per codon can be neglected) an x% difference in nucleotide sequence (corresponding to a 3x% difference in codons) will statistically yield an amino acid difference of 2.3x%. [In the more general case where multiple changes per codon cannot be neglected, computer simulation demonstrates that translation of two DNA sequences which differ from each other by random changes to a given extent (say, 20%) generates deduced amino acid sequences that are much more different (in this case about 40%).] As expected from the genetic code, the nucleotide alignments for orf15, orf48 and orf9 show that mismatches at the
mismatches at the other two positions, and are the major third-base position are much more prevalent than mismatches at the other two positions, and are the major reason for the high amino acid similarity levels.

The conclusion from comparisons between E. coli and S. typhimurium, namely that these open reading frames correspond to real genes, is independently supported by evidence deriving from the individual sequences: (i) the open reading frames were scored as coding regions by the algorithm TESTCODE (Fickett, 1982) (see below) at confidence levels of 96, 96, 88, 95, 99 and 96% for orf15, orf48, orf5, orf9 (and also orf8), respectively; and (ii) they all have strong, well-placed consensus ribosome binding sequences. orf8 also appears to be an authentic gene by these criteria, and has a TESTCODE score of 96%.

The Orf15, Orf48, Orf9 (and also Orf8) sequences were compared against the GenBank database (version 75.0, February 1993) using TFASTA, but no significant similarities were found. In the absence of any clues to the functions of these proteins, we have not attempted to further identify and characterize them, but note briefly below some distinctive characteristics about their deduced amino acid sequences.

Orf15. This sequence has a basic N-terminus followed by a hydrophobic region which is, however, too short to be membrane-spanning. The C-terminus has the sequence TPLP repeated three times in the S. typhimurium sequence and (as APLPTPLP) twice in the E. coli sequence. C-terminal repeats involving T/S and P (but not of the precise form seen in Orf15) have been noted in a variety of DNA binding proteins, including RNA polymerase II and histones (Suzuki, 1990).

Orf48. This sequence is quite basic, with a total of 39 K + R + H residues compared with 17 D + E residues (E. coli data). The N-terminus is basic and is followed by the first of several (seven or eight) hydrophobic regions that are predicted to be membrane-spanning by the algorithm of Engelman et al. (1986). These extend throughout the entire sequence, and suggest that Orf48 is an integral membrane protein with no substantial aqueous domains.

Orf9. No special features were evident for this sequence, which appears to be that of a small soluble protein.

The apparently non-coding region immediately prior to fliE

As was described above, E. coli and S. typhimurium have 2.7 and 0.8 kb of DNA following a series of open reading frames and prior to the first gene of flagellar region IIIb, fliE. The absence of open reading frames of significant size (and also the complete lack of any sequence similarity between the two species) indicated that the DNA was non-coding.
Large non-coding region in the E. coli chromosome

To test this further, we used the algorithm TESTCODE devised by Fickett (1982). This is based on autocorrelation analysis of nucleotides at step 3, and scores the probability that DNA sequence is coding. No knowledge of frame is required, no attention is paid to stop codons, and the algorithm does not depend on signals such as ribosome binding consensus sequences. Typically (e.g. the open reading frames described above), authentic coding sequence scores as such with a confidence level of 90% or higher, whereas non-coding sequence scores very poorly. The algorithm can also be run using a scanning window: in this case, when an authentic reading frame is scanned the confidence level is typically in the 90th percentile most of the time with occasional downward fluctuations; in contrast, 5' and 3' non-coding sequences score as coding at around the 10-20th percentile confidence level, with occasional upward fluctuations.

When the E. coli DNA sequence spanning flagellar region IIIa, the intervening region, and flagellar region IIIb was examined in this way, the results were striking (Fig. 4a). All the way from the flagellar region IIIa genes through orf8 (except for the non-coding regions between adjacent operons) coding probabilities were high; the same was true in flagellar region IIIb. Between orf8 and fliE, however, the coding probabilities were for the most part extremely low; there were a few spikes of higher probability, which might be statistical fluctuations or might represent local regions where coding information has not been entirely lost. A similar situation applied to S. typhimurium (Fig. 4b), but in this case the high probability region corresponding to OrfsE was lacking and the predicted non-coding region was much shorter, as the similarity matrix in Fig. 2 had already revealed.

Further evidence for absence of a coding function for this DNA came from base composition. Whereas coding sequence of DNA in E. coli and S. typhimurium commonly has an A+T content of around 48%, control and other non-coding regions generally have a higher A+T content, often as high as 65% (Haughn et al., 1986). This was found to be true of the 2.7 kb region between orf8 and fliE and the 0.8 kb region between orf9 and fliE, where the mean A+T content was 59% in both cases (data not shown). The later part of the 2.7 kb
region in \textit{E. coli} was especially AT-rich, with a mean content of 66\% over more than 1.4 kb.

In a search of the GenBank database using FASTA, we found no significant similarities to these non-coding sequences; the highest scoring examples (optimized scores in the range 100–150) were AT-rich regions from a variety of sequences.

The DNA in this region lacks any distinguishing characteristics that might provide a clue to its function. It contains some examples of inverted repeats scattered throughout the sequence, but probably not more than would be expected by chance; it does not contain any major direct repeats. No examples of the repetitive extragenic palindromic (REP) sequences that are major direct repeats. No examples of the repetitive content of 66\% over more than 1.4 kb.

The transition from the last flagellar region IIIa operon to \textit{amyA} and the other open reading frames described above is abrupt, with a typically short non-coding region. The same is true between \textit{amyA} and \textit{orf15}, between \textit{orf15} and \textit{orf48}, and in (\textit{E. coli}) between \textit{orf9} and \textit{orf8}. From that point on, all the way to the boundary of flagellar region IIIb, the DNA sequence appears to be non-coding. There are four lines of evidence for a predominantly or exclusively non-coding function, which taken together are compelling: (i) the DNA contains no open reading frames of significant length; (ii) the DNA scores very poorly as coding sequence; (iii) the A+T content is high, a characteristic of non-coding sequence; and (iv) there is no detectable similarity between the \textit{E. coli} and \textit{S. typhimurium} sequences. The existence of substantial amounts of non-coding DNA at the rightmost end of the intervening region is in striking contrast with the extremely tight apposition at the leftmost end, where apparently all redundant DNA between the region IIIa flagellar genes and \textit{amyA} has been squeezed out. Why the two ends of the intervening region should be so different is unclear.

We have examined the literature for other known examples of substantial regions of non-coding DNA on the \textit{E. coli} or \textit{S. typhimurium} chromosomes. Haughn \textit{et al.} (1986) reported a 500 bp gap between two divergent genes, \textit{lep} and an open reading frame of unknown function. More recently, as part of the ongoing \textit{E. coli} genome project, Daniels \textit{et al.} (1992) have described the sequence of 91 kb of DNA between 84.5 and 86.5 min on the chromosome; within that sequence they have found examples of large regions of apparently non-coding DNA, and have coined the phrase 'grey holes' for such regions. Interestingly, in another recent report of an extensive (111 kb) contiguous sequence of the \textit{E. coli} chromosome, from 0 to 24 min, no such grey holes were found (Yura \textit{et al.}, 1992). The largest grey hole described by Daniels \textit{et al.} was 1.2 kb. The one between flagellar regions IIIa and IIIb is, at 2.7 kb, more than twice as large and is to our knowledge the largest that has been found thus far on the \textit{E. coli} chromosome.

It is not known whether grey holes represent sequence that performs a useful function such as, for example, stabilizing higher-order chromosome structure. At least in the case of the grey hole between flagellar regions IIIa and IIIb, it is hard to imagine it having any such function, given its disparate sizes in \textit{E. coli} and \textit{S. typhimurium}.

If these extensive stretches of DNA are non-functional, how might they arise? Genetic rearrangements such as random insertions or deletions will, in general, cause...
disruption or loss of information at the rearrangement boundary, which statistically is likely to be within either a coding region or a control region. If this information is needed for expression of a number of genes (e.g. because it is the promoter of a multi-gene operon), or if loss of one gene product compromises the function of others (e.g. because these proteins form a multi-subunit complex or are components of a multi-enzyme pathway), several kilobases of DNA could easily be rendered useless by the rearrangement. At this point – unless the rearrangement was too deleterious to survive – the sequence would be free to drift.

However, while this scenario can explain the loss of coding content, and also the divergence between the E. coli and S. typhimurium sequences, it does not explain why the useless DNA has been retained. The implication is that the genetic load is small enough that it can be tolerated, much as phenotypically silent mutations are. But, if this is true, it is unclear why large non-coding regions are not commonplace in the bacterial chromosome.

In the case of S. typhimurium, at least, a substantial amount of this DNA has, in fact, been lost (at least two-thirds, if we assume E. coli has lost little or none). It would be interesting in this regard to compare the DNA of a variety of wild-type E. coli strains to see whether this non-coding region differs in length and in sequence and, if so, by how much. It would also be interesting to examine this part of the chromosome in the genus that is most closely related to E. coli, namely, Shigella, which is non-motile but might still have remnants of the flagellar gene system.

Another question that arises is how such a major rearrangement could occur without damage to the flagellar gene system. It does not seem likely that in the ancestral genome there can have been any important flagellar genes between the flID and flIE operons, since there are none now and yet motility has been retained. Yet if these operons were originally adjacent, the disruption must have occurred rather precisely within a non-essential part of sequence in order to leave them both intact. While this might seem a somewhat improbable event statistically, perhaps only such an event would have been able to survive in the face of the selection pressure in favour of motility. Interestingly, the genes in the two regions are now functionally somewhat different, with those in region IIIa being involved in the flagellar filament and other late components of the flagellar assembly process, while those in region IIIb are involved in earlier components such as the switch and the basal-body MS ring. Perhaps this differentiation is in some way related to the events that have caused the physical separation on the chromosome.

We thank Frederick Blattner and Kenneth Rudd for helpful discussions; David DeRosier, Nicholas Ornston and Sandy Parkinson for critically reading this manuscript; and David Morgan for carrying out the database comparisons. This work was supported by USPHS grant AI12202.

References


