Volume 9 Number 7 1981 Nucleic Acids Research

Comparison of the nucleotide sequence of trpA and sequences immediately beyond the trp operon of Klebsiella aerogenes, Salmonella typhimurium and Escherichia coli

Brian P.Nichols*, Miroslav Blumenberg and Charles Yanofsky

Department of Biological Sciences, Stanford University, Stanford, CA 94305, USA

Received 30 December 1980

ABSTRACT

The nucleotide sequence of trpA of Klebsiella aerogenes is presented and compared with the trpA sequences of Salmonella typhimurium and Escherichia coli. The majority of the approximately 200 differences between each pair of trpA's are single nucleotide pair changes that do not alter the amino acid sequence. Codon usage conforms to the general patterns revealed by examination of other prokaryotic gene sequences. However, codon usage in K. aerogenes trpA reflects the high G+C content of the genome of this organism. The DNA sequences just beyond trpA, the presumed transcription termination region, are also compared for the three species. Perusal of these sequences indicates that the secondary structure of the transcript segment just beyond trpA has been preserved, while the primary sequence has diverged appreciably.

INTRODUCTION

Comparisons of nucleotide sequences of several of the genes of the trp operons of enterobacterial genera have revealed that the major feature of the evolutionary process has been silent nucleotide variation (2,13-15). Silent nucleotide changes are those that result in synonymous codon replacements, changes that do not affect the amino acid sequence of the polypeptide product. Analyses of the extent and types of nucleotide differences have suggested that most nucleotide replacements are neutral with respect to gene expression or function.

We report here the nucleotide sequence of trpA of Klebsiella aerogenes, the third enterobacterial trpA sequence so far determined. trpA encodes the α polypeptide subunit of tryptophan synthetase. The amino acid sequence of the α-subunit of K. aerogenes has been described (10). K. aerogenes trpA exhibits a distinctly different, and nonrandom, pattern of nucleotide pair substitutions, when compared to trpA of Salmonella typhimurium and Escherichia coli.

We also present the nucleotide sequence beyond trpA of K aerogenes and S. typhimurium. This region of DNA is believed to contain sequences that are important in the termination of transcription at the end of the trp operon

(18,21-23). Comparison of the sequences reveals several features that are evolutionarily preserved, even though the overall nucleotide sequence is not strictly conserved.

MATERIALS AND METHODS

DNA sequence determinations were performed by the procedure of Maxam and Gilbert (11). Sequences were analyzed in part using the computer programs of Korn et al. (9).

Restriction endonucleases were purchased from either New England Bio Labs or Bethesda Research Laboratory, or prepared in this laboratory by published procedures. Restriction endonuclease incubation conditions were those recommended by commercial suppliers. The construction of plasmid pKA2 will be described elsewhere (Blumenberg & Yanofsky, manuscript in preparation). DNA was isolated as previously described (8).

RESULTS

Nucleotide Sequence Determination. Plasmid pKA2 contains 10.6 kb of *Klebsiella aerogenes* DNA inserted in the *Bam*HI site of pBR322. The *K. aerogenes* DNA insert is composed of two *Bam*HI fragments, 3.4 and 7.2 kb in length, and contains the entire *trp* operon. Genetic complementation tests indicated that the *Bam*HI site is within *trpA*. Further analyses enabled us to construct the restriction map shown in the upper portion of Fig. 1.

The sequencing strategy used is also illustrated in Fig. 1. With the exception of the *Sst*II-*Bam*HI and the *Bam*HI-*Sal*I fragments, all the restriction fragments indicated were derived from the 1080 bp *Sma*I fragment. The entire sequence of both strands of *trpA* was determined, and, with the exception of the *Bam*HI site, all restriction sites were overlapped. Colinearity of the *K. aerogenes trpA* nucleotide sequence with the reported amino acid sequence (10) and with the nucleotide and amino acid sequences of *E. coli* and *S. typhimurium* (13) is convincing evidence that no segment of the DNA sequence around the *Bam*HI site was omitted.

The polypeptide product of *K. aerogenes trpA*, the α-subunit of tryptophan synthetase, is 269 amino acid residues long, containing one extra residue at the carboxyl terminus relative to the α-subunits of *S. typhimurium* or *E. coli*. As in these species, the α-subunit of *K. aerogenes* does not contain tryptophan. The predicted amino acid sequence agrees quite closely with that determined by Li & Yanofsky (10). Figure 2 shows the nucleotide and deduced amino acid sequence of *K. aerogenes trpA* and its polypeptide product and il-

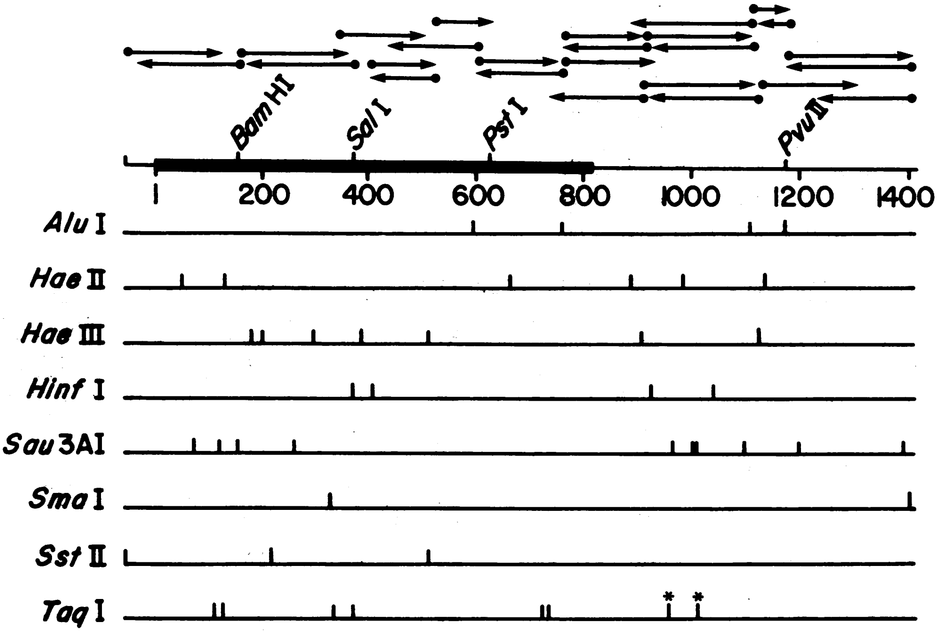

Fig. 1. Nucleotide sequence determinations and restriction sites used to establish the K. aerogenes trpA sequence. Arrows indicate the extent of each sequence determination. The asterisks over two TaqI restriction sites indicate predicted, but uncleaved sites. The sequences are protected from cleavage because they overlap methylated Sau3AI recognition sites.

lustrates the differences from the corresponding genes and proteins of E. coli and S. typhimurium.

Sequence Divergence of trpA and its Polypeptide Product. The trpA nucleotide sequences of K. aerogenes, E. coli, and S. typhimurium are nearly equally divergent from each other. The K. aerogenes sequence differs from that of E. coli and S. typhimurium at 189 (23.5%) and 201 (25.0%) nucleotides, respectively. E. coli and S. typhimurium trpA's differ by 199 (24.6%) nucletides. The amino acid sequences of the trpA polypeptides are also equally divergent from one another. The K. aerogenes chain differs from those of E. coli and S. typhimurium by 34 amino acid residues (12.7%) and 42 amino acid residues (16.0%), respectively, while E. coli and S. typhimurium trpA polypeptides differ by 40 amino acid residues (14.9%). The deduced evolutionary braneching order was determined by the method of Fitch and Margoliash (4), and is presented in Figure 3. The data suggest that K. aerogenes and E. coli

```
KLEBSIELLA AEROGENES                    ARG GLY ASP LYS ASP ILE PHE THR VAL HIS ASP ILE LEU LYS ALA ARG GLY GLU ILE END
                                        CGC GGC GAC AAA GAC ATT TTC ACC GTA CAT GAC ATC CTG AAA GCG CGA GGG GAA ATC TG
ESCHERICHIA COLI                                  T           C           T   C   T   T T         A
SALMONELLA TYPHIMURIUM                        A   T           C   T           C   T
                                     10                                      20                                      30
MET GLU ARG TYR GLU THR LEU PHE ALA GLN LEU LYS ASN ARG GLN GLU GLY ALA PHE VAL PRO PHE VAL THR LEU GLY ASP PRO GLY PRO
ATG GAA CGC TAT GAG ACG CTG TTT GCC CAA CTG AAA AAT CGC CAG GAA GGC GCC TTC GTC CCC TTT GTC ACC CTT GGC GAT CCG GGA CCG
                    SER                         GLU     LYS                                                         ILE
              C   A T T               G T     G G G     A A           A       T   T   C       G   C   T           C ATT
                    ASN                     ASN ASP     ARG                                                         ILE
              C   A  AT T A               C   C G        G            T   T           C   G       G       C   T   C ATT
                                     40                                      50                                      60
GLU GLN SER LEU LYS ILE ILE ASP ALA LEU ILE GLU GLY GLY ALA ASP ALA LEU GLU LEU GLY ILE PRO PHE SER ASP PRO LEU ALA ASP
GAG CAG TCG CTG AAA ATT ATC GAT GCG CTG ATC GAA GGC GGC GCC GAT GCC CTT GAA CTG GGG ATC CCT TTC TCC GAC CCG CTG GCG GAC
                                THR             ALA
          A T                   A     A   T      C    T   T   C   G   G   G T A   T       C               A           T
                                THR         ASP ALA                                 VAL
  A       A               T   C A A       T   T  C            C   T   G             G T   C           T           C   T
                                     70                                      80                                      90
GLY PRO THR ILE GLN GLY ALA ALA LEU ARG ALA PHE ALA ALA GLY VAL THR PRO ALA GLN CYS PHE GLU MET LEU ALA ALA ILE ARG GLN
GGC CCG ACC ATC CAG GGC GCG GCC CTG CGC GCC TTC GCC GCG GGA GTG ACC CCG GCG CAG TGC TTT GAG ATG CTG GCG GCG ATC CGC CAG
                    ASN     THR                                                                         LEU
          G   T   A AA    C A T               T   G   A   T       T       A   A   T       A           A CT    T
                    ASN     ASN                                                                         LEU         GLU
      T             AAT     AA  T A                   T   C   C   G       T       T       A             CT    T   T G A
                                    100                                     110                                     120
LYS HIS PRO THR ILE PRO ILE GLY LEU LEU MET TYR ALA ASN LEU VAL PHE SER PRO GLY ILE ASP ALA PHE TYR ALA GLN CYS ALA ARG
AAG CAT CCG ACC ATC CCA ATC GGC CTG CTG ATG TAT GCC AAC CTC GTC TTC AGC CCG GGC ATC GAT GCG TTC TAT GCC CAG TGC GCC CGC
                                                                    ASN LYS             GLU                     GLU LYS
  A   C           T   C   T         T                 T   G   G   T  A  AAA       T      A    T                  AA AAA
                                                                    ASN ASN                             ARG     GLU GLN
  A   C           T   G   T           A       C   G   T   G   G      AT AAC       A                      GT   T  AA  AG
                                    130                                     140                                     150
VAL GLY VAL ASP SER VAL LEU VAL ALA ASP VAL PRO VAL GLU GLU SER ALA PRO PHE ARG GLN ALA ALA MET ARG HIS ASN ILE ALA PRO
GTC GGC GTC GAC TCG GTG CTG GTG GCC GAC GTG CCG GTG GAA GAG TCC GCC CCG TTC CGC CAG GCG GCG ATG CGC CAT AAC ATC GCG CCG
                                                                                            LEU             VAL
              T               T       T       A   T               G   C               C     T     T   C     G     A   T
                                                                                            LEU
  T       A   T   C           C   A   T   C       T       A   G       C               A     T A   G       T
                                    160                                     170                                     180
ILE PHE ILE CYS PRO PRO ASN ALA ASP ASP ASP LEU LEU ARG GLN ILE ALA SER TYR GLY ARG GLY TYR THR TYR LEU LEU SER ARG ALA
ATT TTC ATC TGC CCG CCC AAT GCG GAT GAC GAT TTA CTG CGC CAG ATT GCC TCT TAT GGC CGC GGC TAC ACC TAT CTG CTG TCG CGC GCC

  C                   A       C           C C G               A           C   T   T   T             T         A   A   A
                                                            VAL                                                     SER
  C                   A                     C T             G C   A       C           T           C       T       T T G
                                    190                                     200                                     210
GLY VAL THR GLY ALA GLU ASN ARG ALA ALA LEU PRO LEU HIS HIS LEU VAL GLU LYS LEU ALA GLU TYR HIS ALA ALA PRO PRO LEU GLN
GGA GTG ACG GGT GCG GAA AAT CGC GCC GCG CTG CCG CTG CAC CAC CTG GTG GAG AAG CTG GCG GAA TAT CAC GCC GCG CCG CCG CTG CAG
                                                    ASN             ALA         LYS         ASN
  C       C   C   A       C             T A   C   C A T   T       T  C          AAA   G   C A     T   A   T   A T
                                GLY                             ILE             LYS                         ALA
  T   C   C   C       C   T  G    A T       T     T   T   C A T           T AAA   G   C   T           T G   T A
                                    220                                     230                                     240
GLY PHE GLY ILE SER ALA PRO GLU GLN VAL SER ALA ALA ILE ASP ALA GLY ALA ALA GLY ALA ILE SER GLY SER ALA ILE VAL LYS ILE
GGC TTT GGT ATC TCC GCG CCG GAG CAG GTG AGC GCC GCC ATT GAC GCC GGA GCC GCC GGG GCT ATC TCC GGT TCC GCC ATC GTC AAA ATC
                            ASP         LYS
  A           T       C       T       A  AA   A   G       T   A       T   G   C   G   T   T       G       T   T
                    SER                                 VAL ARG
      C             T         A         TCT   G     G G CGT         G   G   T   C               C   A       T       G   T
                                    250                                     260                                 269
ILE GLU ARG HIS LEU ASP GLU PRO GLN THR MET LEU ASP GLU LEU LYS ALA PHE VAL GLN SER LEU LYS ALA ALA THR LYS THR ALA END
ATC GAA CGC CAC CTC GAT GAG CCG CAG ACC ATG CTG GAC GAG CTG AAA GCC TTC GTC CAG AGC CTG AAG GCA GCG ACC AAA ACC GCC TGA
        GLN     ILE ASN         GLU LYS         ALA ALA         VAL             PRO MET                 ARG SER END
      G  AA   T A T A         A G    AA          CG  CA          TT   T   A   A CCG A     A   G       G CGC  GT TAA
        LYS ASN     ALA SER     LYS GLN         ALA         ARG SER         SER ALA MET             SER ARG ALA END
      G AAA A        CG TCT   C A A CAG     T    CG       C  GG T     T     TCA GC  A     A   C   C  G  CGC G A TAA
```

Fig. 2. The nucleotide and amino acid sequences of three *trpA* genes. The complete nucleotide and amino acid sequences of *K. aerogenes trpA* is given. Only sequence differences that occur in *E. coli* and *S. typhimurium trpA* are listed below the *K. aerogenes* sequence.

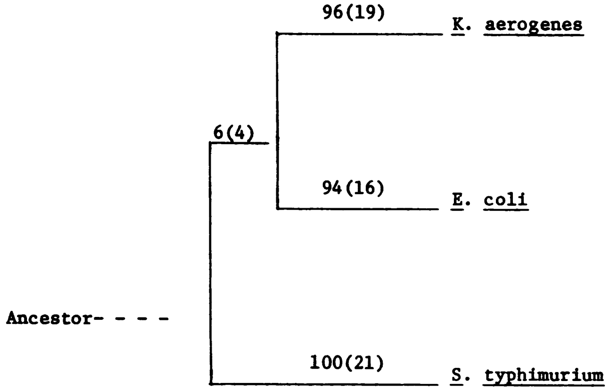

Fig. 3. Evolutionary branching order of the three trpA genes. Calculated nucleotide and amino acid (parentheses) replacements are shown for each branch.

trpA diverged from one another slightly after S. typhimurium trpA diverged from a common ancestral sequence. Previous branching orders of these three species, calculated from immunological studies of different enzymes, suggested that the three organisms diverged from a common ancestor at about the same time (2). The use of nucleotide sequences to determine the branching order allows a more detailed analysis of evolutionary relationships.

An hypothetical "ancestral" trpA nucleotide sequence can be constructed from the nucleotide sequences of the present day organisms, assuming maximum parsimony of nucleotide and amino acid sequences. If two of three sequences contain a particular nucleotide at one position, it is assumed that the nucleotide belongs to the ancestral sequence. The use of only three beginning sequences results in an "average" ancestral sequence. Furthermore, the method cannot distinguish substitutions that have arisen between the common ancestor and S. typhimurium from those that have arisen between the common ancestor and the branch point of E. coli and K. aerogenes. Substitutions in the latter branch are assigned as having arisen between the ancestor and S. typhimurium. However, since the number of substitutions between the ancestor and the E. coli-K. aerogenes branch point is small compared to the number between the ancestor and S. typhimurium (6 vs. 100), the error in subsequent analysis of nucleotide substitutions is not greatly affected. All but 33 nucleotides in the ancestral sequence can be assigned.

A comparison of the ancestral and descendant nucleotide sequences indicates the type of substitutions that may have occurred in trpA during the divergence from a common ancestor, and these are shown in Table 1. A total of 247 substitutions can be analyzed. Of these, 58% are transitions. The four types of transversion substitutions represent 42% of the total substitutions. Each type of transversion has occurred with nearly equal frequency, with the notable exception of A:T→T:A. A similar pattern of nucleotide substitutions has been observed in the divergence of trp(G)D nucleotide sequences (14).

In the individual lineages, the particular types of substitutions reflect primarily the evolution of the individual genomes toward their contemporary G+C content (Table 2). For example, the *K. aerogenes* genome has evolved to approximately 56% G+C, whereas *E. coli* and *S. typhimurium* have evolved to 51% and 52% G+C, respectively (19). Changes in G+C content reflecting the higher genomic G+C content of *K. aerogenes* can be seen in the total coding sequence, first and second positions of codons, and third positions of codons. The most dramatic difference in G+C content, however, is in codon third positions, which of course have greater freedom of substitution, due to the degeneracy of the genetic code.

The longest segment of the nucleotide sequence that is entirely conserved in all three species is a region of 25 bp which occurs around the trpB-trpA junction (Fig. 2). This region includes the ribosome binding site

Table 1
Nucleotide Substitutions in trpA

Substitution	Ancestor ↓ *K. aerogenes*	Ancestor ↓ *E. coli*	Ancestor ↓ *S. typhimurium*	Total
Transitions				
G:C → A:T	15	43	30	88
A:T → G:C	34	5	17	56
Transversions				
G:C → C:G	8	13	17	38
G:C → T:A	5	11	17	33
A:T → T:A	1	1	2	4
A:T → C:G	18	4	6	28

Table 2
G+C Content of trpA's

	% G+C		
	K.a.	S.t.	E.c.
genome	56	52	51
trpA coding sequence	64.8	57.3	54.0
trpA codon 3rd positions	83.3	63.4	56.3
trpA codon 1st + 2nd positions	55.6	54.2	52.8

K.a., *Klebsiella aerogenes*; S.t., *Salmonella typhimurium*; E.c., *Escherichia coli*

as well as the specialized intercistronic sequence -TGATG-. This 5 bp sequence contains overlapping translation termination and initiation codons, and may be involved in coordination of the synthesis of the α- and β- polypeptides of tryptophan synthetase (16).

Codon Usage in trpA. The most outstanding feature of codon usage in *K. aerogenes* trpA is the very high frequency (83%) of codons ending in either G or C (Table 3). The bias occurs primarily in codon families that are 3- or 4-fold degenerate. In those codon families where 3- or 4-fold degeneracy can occur, there is a gradient of increasing G+C content from *E. coli* to *S. typhymurium* to *K. aerogenes* (see, for example, the Val, Ala and Ile families), and this gradient follows the G+C content of the organisms' genomes. *Serratia marcescens*, another enterobacterial organism with high G+C content (59%) has a similar bias toward G+C in codon usage in trpG (14).

The trends in codon usage that have been observed in the trp operon structural genes of *E. coli* and *S. typhimurium* (3,13-15) are also evident in *K. aerogenes* trpA. For example, CTG(Leu), GGPy(Gly), and CGPy(Arg) are most frequently used, while ATA(Ile), CGPu(Arg) and AGPu(Arg) are rarely used codons. In this regard, the several species of Enterobacteriacae so far studied have a common pattern of rarely and frequently used codons. Beyond this, other major fluctuations in codon usage reflect the G+C content of the organisms.

Comparison of Sequences Beyond trpA. The nucleotide sequences of approximately 90 bp beyond the end of each trpA gene are shown in Figure 4.

Table 3
Codon Frequencies in trpA's

	K.a.	S.t.	E.c.		K.a.	S.t.	E.c.		K.a.	S.t.	E.c.		K.a.	S.t.	E.c.
Phe TTT	4	4	7	Ser TCT	1	3	3	Tyr TAT	6	1	3	Cys TGT	0	2	1
Phe TTC	8	8	5	Ser TCC	5	5	3	Tyr TAC	1	6	4	Cys TGC	3	1	2
Leu TTA	1	4	2	Ser TCA	0	3	2	End TAA	0	1	1	End TGA	1	0	0
Leu TTG	0	3	6	Ser TCG	3	4	2	End TAG	0	0	0	Trp TGG	0	0	0
Leu CTT	2	3	0	Pro CCT	1	3	3	His CAT	2	4	2	Arg CGT	0	5	2
Leu CTC	2	4	2	Pro CCC	2	4	4	His CAC	4	1	2	Arg CGC	12	7	8
Leu CTA	0	1	1	Pro CCA	1	1	5	Gln CAA	1	1	4	Arg CGA	0	0	1
Leu CTG	22	13	16	Pro CCG	16	9	7	Gln CAG	12	9	8	Arg CGG	0	2	0
Ile ATT	4	10	13	Thr ACT	0	0	2	Asn AAT	3	6	4	Ser AGT	0	0	1
Ile ATC	15	7	6	Thr ACC	8	5	3	Asn AAC	2	5	5	Ser AGC	3	1	0
Ile ATA	0	1	1	Thr ACA	0	1	0	Lys AAA	5	6	11	Arg AGA	0	0	0
Met ATG	5	5	5	Thr ACG	2	1	4	Lys AAG	3	2	2	Arg AGG	0	1	0
Val GTT	0	3	6	Ala GCT	1	6	3	Asp GAT	7	9	9	Gly GGT	3	3	8
Val GTC	7	8	4	Ala GCC	24	15	13	Asp GAC	7	4	4	Gly GGC	12	15	9
Val GTA	0	1	2	Ala GCA	1	5	11	Glu GAA	8	12	8	Gly GGA	4	0	2
Val GTG	8	5	5	Ala GCG	18	14	13	Glu GAG	8	4	9	Gly GGG	2	2	0

K.a., *Klebsiella aerogenes*; S.t., *Salmonella typhimurium*; E.c., *Escherichia coli*

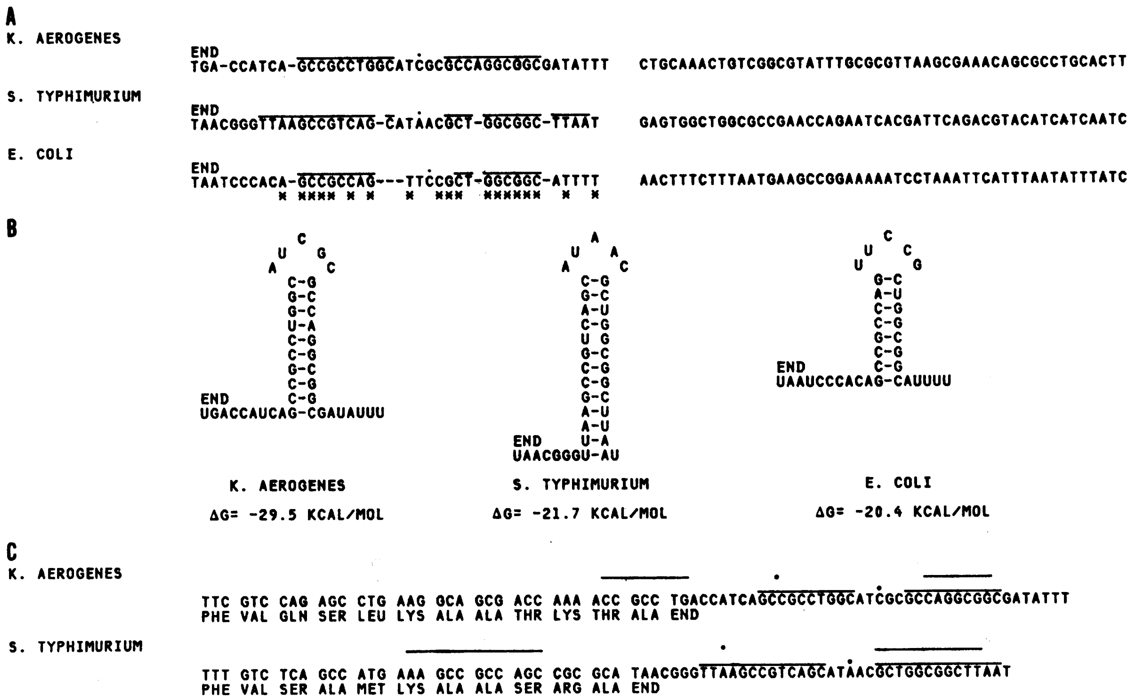
A
K. AEROGENES
END
TGA-CCATCA-GCCGCCTGGCATCGCGCCAGGCGGCGATATTT CTGCAAACTGTCGGCGTATTTGCGCGTTAAGCGAAACAGCGCCTGCACTT
S. TYPHIMURIUM
END
TAACGGGTTAAGCCGTCAG-CATAACGCT-GGCGGC-TTAAT GAGTGGCTGGCGCCGAACCAGAATCACGATTCAGACGTACATCATCAATC
E. COLI
END
TAATCCCACA-GCCGCCAG---TTCCGCT-GGCGGC-ATTTT AACTTTCTTTAATGAAGCCGGAAAAATCCTAAATTCATTTAATATTTATC
B
K. AEROGENES
ΔG= -29.5 KCAL/MOL
S. TYPHIMURIUM
ΔG= -21.7 KCAL/MOL
E. COLI
ΔG= -20.4 KCAL/MOL
END
UGACCAUCAG-CGAUAUUU
UAACGGGU-AU
UAAUCCCACAG-CAUUUU
C
K. AEROGENES
TTC GTC CAG AGC CTG AAG GCA GCG ACC AAA ACC GCC TGACCATCAGCCGCCTGGCATCGCGCCAGGCGGCGATATTT
PHE VAL GLN SER LEU LYS ALA ALA THR LYS THR ALA END
S. TYPHIMURIUM
TTT GTC TCA GCC ATG AAA GCC GCC AGC CGC GCA TAACGGGTTAAGCCGTCAGCATAACGCTGGCGGCTTAAT
PHE VAL SER ALA MET LYS ALA ALA SER ARG ALA END

Fig. 4. a) Nucleotide sequences beyond trpA, aligned for maximum nucleotide sequence identity. The trpA translation termination codon is indicated by END. Dots denote the axes of dyad symmetry and lines indicate the extent of symmetry. The break in the sequence denotes the presumed terminus of trpmRNA and is based on analogy with the E. coli sequence. Asterisks mark nucleotide sequence identity in the three organisms. b) Proposed secondary structure at the end of trpmRNA. ΔG values were calculated by the rules of Tinoco et al. (20) and Borer et al. (1). c) Illustration of two overlapping regions of dyad symmetry at the termini of K. aerogenes and S. typhimurium trpA. Lines immediately above the sequence represent the symmetry elements illustrated in a) and b). Lines above these indicate the limits of the potential competing symmetry elements.

Our sequence of this region of E. coli DNA is in complete agreement with the previously determined sequence (21). Each sequence shows the structural features believed to be important in transcription termination: a G+C-rich region followed by an A+T-rich sequence and a region of dyad symmetry that may form a hairpin structure in the mRNA (Fig. 4b). E. coli trp mRNA produced in vivo has a 3' terminus at the end of the U-rich region (21) (indicated by the break in the sequence). The in vivo end points of the K. aerogenes and S. typhimurium trp mRNA's have not been determined; for comparative purposes we presume them to be at the sites indicated in Figure 4a.

The structural features mentioned are preserved in all three organisms even though the nucleotide sequence itself is not highly conserved. Several insertions and deletions, in addition to nucleotide substitutions, are found when comparing the sequences. The most highly conserved feature of the nucleotide sequence is in the G+C-rich region within the "stem" of the potential mRNA secondary structure. Each of the structures varies in length and stability, however. Both K. aerogenes and S. typhimurium have longer and slightly more stable secondary structures than E. coli.

One additonal conserved feature of these sequences is the presence of an alternate, and possibly competing, secondary structure in the vicinity of the translation stop codon (Fig. 4c). Wu and Platt (21) observed that in E. coli an 8 bp sequence within the trpA structural gene was complementary to the proximal portion of the hairpin loop structure at the end of the message. Such complementarity could allow the formation of a secondary structure in the mRNA other than that shown in Figure 4b. Both the K. aerogenes and S. typhimurium sequences show a similar alternate secondary structure. However, in these cases the sequence within trpA is complementary to the distal portion of the hairpin loop structure. Translation of trpA mRNA could preclude the formation of these alternate structures and ensure the formation of the hairpin loop structure at the extreme terminus of the mRNA. The predicted stabilities (1,20) of the alternate structures (K. aerogenes, $\Delta G = -13.4$ Kcal/mol; S. typhimurium, $\Delta G = -27.0$ Kcal/mol; E. coli, $\Delta G = -16.4$ Kcal/mol) along with the fact that the structures are conserved in all three organisms, suggest that they have some as yet undetermined functional significance.

DISCUSSION

A comparison of homologous nucleotide sequences among related organisms provides a framework for assessing some of the molecular events of evolution. In previous studies of trp gene sequences (3,13-15) it was observed that the majority of nucleotide differences within the coding region of the trp operon are silent and occur in codon third positions (14). This pattern is also seen in the comparisons of the trpA sequences of K. aerogenes with those of S. typhimurium and E. coli (Figure 3). The longest trpA nucleotide sequence identical in all three bacteria is the 25 bp region spanning the trpB/trpA junction. The nucleotide sequence identity implies an important role in translation initiation. The least conserved region is the carboxy-terminal portion of the protein where only 9 of 27 amino acid residues are the same in all three species. This is compared with 207 identical amino acid residues in the remaining 242 residues. It is probable that this portion of the polypeptide is not critical for catalytic function or for subunit interactions and is therefore subject to greater variation.

Comparisons of codon usage in the three trpA genes supports the hypothesis of Grantham et al. (5,6) that codon use in bacteria is a genomic strategy modulated for gene expressivity. In other words, genes that are expressed intermittently or at moderate levels show a pattern of codon use that reflects the organism's genomic G and C content. The trpA genes fall into this category, and the relatively high G+C content of K. aerogenes DNA (56%) is reflected in the extremely high use of G and C in codon third positions (83%). Genes that are expressed at very high levels, such as ribosomal protein genes and major membrane protein genes exhibit a more restricted pattern of codon use, based probably on avoidance of codons of the strongest binding energies (6). A particularly dramatic example of these two influences operating in the same organism is found in a comparison of Serratia marcescens lpp and trpG. S. marcescens has a genomic G+C content of 59% (19). The third positions of the trpG codons contain 82% G and C (14), similar to that observed in K. aerogenes trpA. However, codon usage in the highly expressed lpp gene of S. marcescens (12) closely resembles that observed in the highly expressed E. coli genes (third position G and C = 53%). As a result, some codon families (e.g., Val, Ala) exhibit a complete reversal of codon preference from codons ending in G and C (trpG) to codons ending in A and T (lpp).

A comparison of nucleotide sequences beyond the end of trpA reveals features that are preserved among the three organisms, although the nucleotide sequences are not strictly conserved. A G+C-rich stem-and-loop structure, followed by an A+T-rich sequence, in addition to mutually exclusive alternate secondary structures, are features common to all three sequences. These features had been thought to be involved in rho-dependent transcription termination at the end of the trp operon (17,21). However, recent evidence suggests that sequences required for the termination event occur distal to these structures; i.e., it has been demonstrated that in E. coli a sequence 60 to 100 bp beyond the end of trpA is involved in efficient termination in vivo (7,22) and in vitro (23). We have observed that in S. typhimurium, as well, efficient termination in vivo requires the presence of nucleotide sequences between 75 and 200 bp distal to the end of trpA (data not shown). In addition, our in vitro studies indicate that rho-dependent transcription termination occurs at several sites within this distal region (Nichols & Yanofsky, unpublished). Although some termination may occur at the site indicated in Figure 4, it is not the major site of termination in vitro and is not rho-dependent. Wu, Christie & Platt (23) have similarly demonstrated the existence of a strong in vitro rho-dependent termination site about 250 bp beyond the presumed site of termination in vivo. These data raise the question of the precise function of the structures observed near the trpA terminus. One possibility is that the stem-and-loop structure impedes exonucleolytic degradation from the 3' terminus of the mRNA. The observed 3' terminus of trp mRNA (21) might be the steady-state product of distally terminated transcription followed by degradation of sequences not protected by translation or the presence of a secondary structure. If this were the case, then transcription must in fact terminate at a site, or sites, well beyond the major region of secondary structure, past the apparent 3' terminus of in vivo isolated trp mRNA.

ACKNOWLEDGMENTS

We thank our colleagues for helpful discussions and critical review of this manuscript. The authors appreciate the exchange of preprints with Dr. Terry Platt. B.P.N. was an American Cancer Society postdoctoral fellow. M.B. is a Career Investigator Fellow of the American Heart Association. C.Y. is a Career Investigator of the American Heart Association.

These studies were supported by grants from the United States Public Health Service (GM09738), the National Science Foundation (PCM77-24333), and the American Heart Association (69C-15).

*Present address: Department of Biological Sciences, University of Illinois at Chicago Circle,
Box 4348, Chicago, IL 60680, USA

REFERENCES

1. Borer, P. N., Dengler, B., Tinoco, I., Jr., and Uhlenbeck, O. C. (1974) J. Mol. Biol. 85, 843-853.

2. Cocks, G. T., and Wilson, A. C. (1972) J. Bacteriol. 110, 793-802.
3. Crawford, I. P., Nichols, B. P., and Yanofsky, C. (1980) J. Mol. Biol. 142, 489-502.
4. Fitch, W. M., and Margoliash, E. (1967) Science 155, 279-284
5. Grantham, R., Gautier, C., and Gouy, M. (1980) Nucl. Acids Res. 8, 1893-1912
6. Grantham, R., Gautier, C., Gouy, M., and Mercier, R. (1981) Nucl. Acids Res. 9, r43-r74
7. Guarente, L. P., Beckwith, J., Wu, A. M., and Platt, T. (1969) J. Mol. Biol. 133, 189-197
8. Gunsalus, R. P., Zurawski, G., and Yanofsky, C. (1979) J. Bacteriol. 140, 106-133
9. Korn, L. J., Queen, C. L., and Wegman, M. N. (1977) Proc. Natl. Acad. Sci. 74, 4401-4405
10. Li, S., and Yanofsky, C. (1973) J. Biol. Chem. 248, 1837-1843
11. Maxam, A. M., and Gilbert, W. (1977) Proc. Natl. Acad. Sci. 74, 560-564
12. Nakamura, K., and Inouye, M. (1980) Proc. Natl. Acad. Sci. 77, 1369-1373
13. Nichols, B. P., and Yanofsky, C. (1979) Proc. Natl. Acad. Sci. 76, 5244-5248
14. Nichols, B. P., Miozzari, G. F., van Cleemput, M., Bennett, G. N., and Yanofsky, C. (1980) J. Mol. Biol. 142, 503-517
15. Nichols, B. P., van Cleemput, M., and Yanofsky, C. (1981) J. Mol. Biol., in press
16. Oppenheim, D., and Yanofsky, C. (1980) Genetics 95, 785-795
17. Platt, T. (1980) in The Operon, Miller, J. and Reznihoff, W. S., Eds., pp. 262-302, Cold Spring Harbor
18. Rose, J. K., and Yanofsky, C. (1971) J. Bacteriol. 108, 615-618
19. Sober, H. A. (Ed.) (1970) Handbook of Biochemistry: Selected Data for Molecular Biology. CRC Press, West Palm Beach, Florida. 2nd ed., pp. H83-H85
20. Tinoco, I., Jr., Borer, P. N., Dengler, B., Levine, M. D., Uhlenbeck, O. C., Crothers, D. M., and Gralla, J. (1973) Nature New Biol. (London) 246, 40-41
21. Wu, A. M., and Platt, T. (1978) Proc. Natl. Acad. Sci. 75, 5442-5446
22. Wu, A. M., Chapman, A. B., Platt, T., Guarante, L. P., and Beckwith, J. (1980) Cell 19, 829-836
23. Wu, A. M., Christie, G. E., and Platt, T. (1981) Proc. Natl. Acad. Sci., submitted