Volume 9 Number 2 1981 Nucleic Acids Research

Nucleotide sequence of the thrB gene of E. coli, and its two adjacent regions; the thrAB and thrBC junctions

Pascale Cossart, Michaël Katinka and Moshe Yaniv

Département de Biologie moléculaire, Institut Pasteur, 25, rue du Dr. Roux, 75015 Paris, France

Received 19 November 1980

ABSTRACT.

We have sequenced a DNA fragment containing the Escherichia coli thrA-thrB junction, the complete thrB gene and the thrB-thrC junction. The intergenic sequence thrA and thrB is only one base pair. The coding region for homoserine kinase is 927 base pairs long. It is followed by 114 base pair segment in an open reading frame predicting that thrC begins just after the non-sense codon of thrB. The presence at the end of thrA and of thrB of sequences that can pair with the 3' end of the 16 S ribosomal RNA suggests that reinitiation of translation occurs at the end of the two genes. The deduced aminoacid sequence for homoserine kinase shows no striking homology with aspartokinase I homoserine dehydrogenase I.

INTRODUCTION.

Recent developments of rapid methods for DNA sequence determination have given essential information on the signals for initiation and termination of transcription. Moreover, one can precisely locate the genes between their translational start and stop signals on the corresponding mRNA. On the other hand, DNA sequence determination is now a rapid and elegant way of determinating protein primary structure. Comparison of both DNA and protein sequences is now a tool of choice for the study of evolutionary processes.

To learn more about the regulation of transcription and translation of the threonine operon (thrABC) as well as to compare the three different gene products of this biosynthetic operon, we have identified the thrA-thrB junction, determined the complete nucleotide sequence of the thrB gene, and the 114 base pairs which follow it. Ribosomal binding sites present at the ends of thrA and thrB, predict that reinitiation of translation

can occur on the polycistronic mRNA. Comparison by extensive computer analysis of the two first genes of the threonine operon, as well as that of their translational products, did not reveal any extensive homology.

MATERIALS AND METHODS.

a) Molecular cloning of the thrB genes.

The plasmid pIPII, carrying the thrA and the thrB genes between the Hind III and Eco RI sites of pBR322 was used for the study described here. Construction of the hybrid plasmid, its purification as well as the isolation of restriction fragments were as previously described (1).

b) DNA sequencing.

The procedures of Maxam et Gilbert (2) and Sanger et al.(3) were used. Labeling of the 5' ends of DNA fragments with (γ-^{32}P)-ATP (3000 Ci/mmole, Amersham) and T_4 polynucleotide kinase was done by the exchange reaction of Berkner et al. (4). The sequences determined by the chain terminator technique (3) were obtained after randomly cloning a Sau 3A digest of the Eco RI site containing Hind II fragment of pIPII, in the single stranded phage vector M13mp2/Bam (5). The primer used was a 96 base pairs Eco RI fragment from phage M13mp2962 (6).

Sequencing acrylamide urea gels, at the beginning of this work, were made and run as originally described (2) and then were the thin gels of Sanger et al. (7).

c) Computer analysis.

Analysis of the nucleotide sequence was done with the programs of R. Staden (8,9,10), and F. Schaeffer (manuscript in preparation). A two dimensional dot matrix comparison program was developed by P. Herbomel (personal communication) in which the two genes or the two proteins are compared one to another, base by base or aminoacid by aminoacid. Prediction of the protein secondary structure was done according to Garnier et al. (11). The computer facilities of the Pasteur Institute (Unité Calcul) were used for most of these studies.

RESULTS.

a) Sequence of the thrB gene.

In a previous study, we showed that a Hind III - Eco RI fragment from a λdthr transducing phage cloned in pBR322 contained the thrA and thrB genes of E.coli (1). The complete nucleotide sequence of thrA was determined (12). To locate precisely the thrB gene, we sequenced towards thrB, starting from the Hinf I site located at the very end of thrA. The sequence strategy is represented on figure 1. The sequence of the thrB gene and its two adjacent regions towards thrA and thrC is shown on figure 2.

The gene was sequenced for over 80% of both strands. It is 927 base pairs and codes for a protein which is 309 aminoacids long. The predicted N terminal sequence agrees with that determined by protein sequencing (13) except for the N terminal methionine which has been removed in the mature protein, as it is often the case in E.coli.

The codon usage, as shown in Table 1, is not random. One feature of this repartition is the net preference for a codon corresponding to the major tRNA species: CUG for Leu (14), GGY (Y= py-

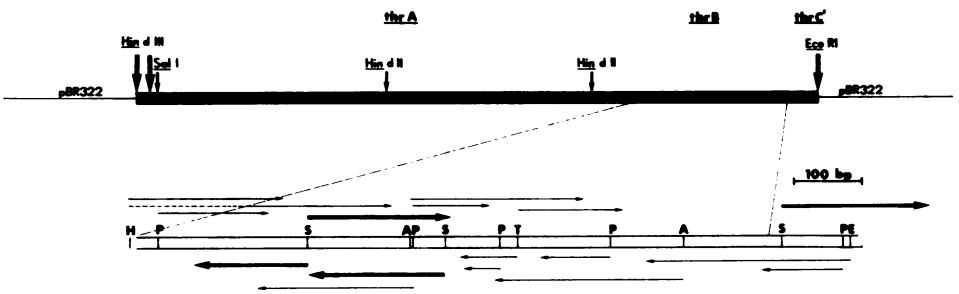

Figure 1. The pBR322 thr (pIPII) hybrid plasmid and the sequencing strategy of the thrB gene.

A/ A restriction map with the Hind III and Eco RI sites used for the cloning of the 4 kbp thr fragment containing the thrA, thrB, and part of the thrC (thrC') genes.

B/ The sequencing strategy of the thrB gene. The arrows indicating the sites used for 5' labeling as well as the direction and extent of the sequences are determined (E = Eco RI ; H = Hinf I, P = Hpa II, S = Sau 3A, A = Hae III, T = Tag I).

The thick arrows (➡) are the sequences determined with the dideoxynucleotide-terminator technique.

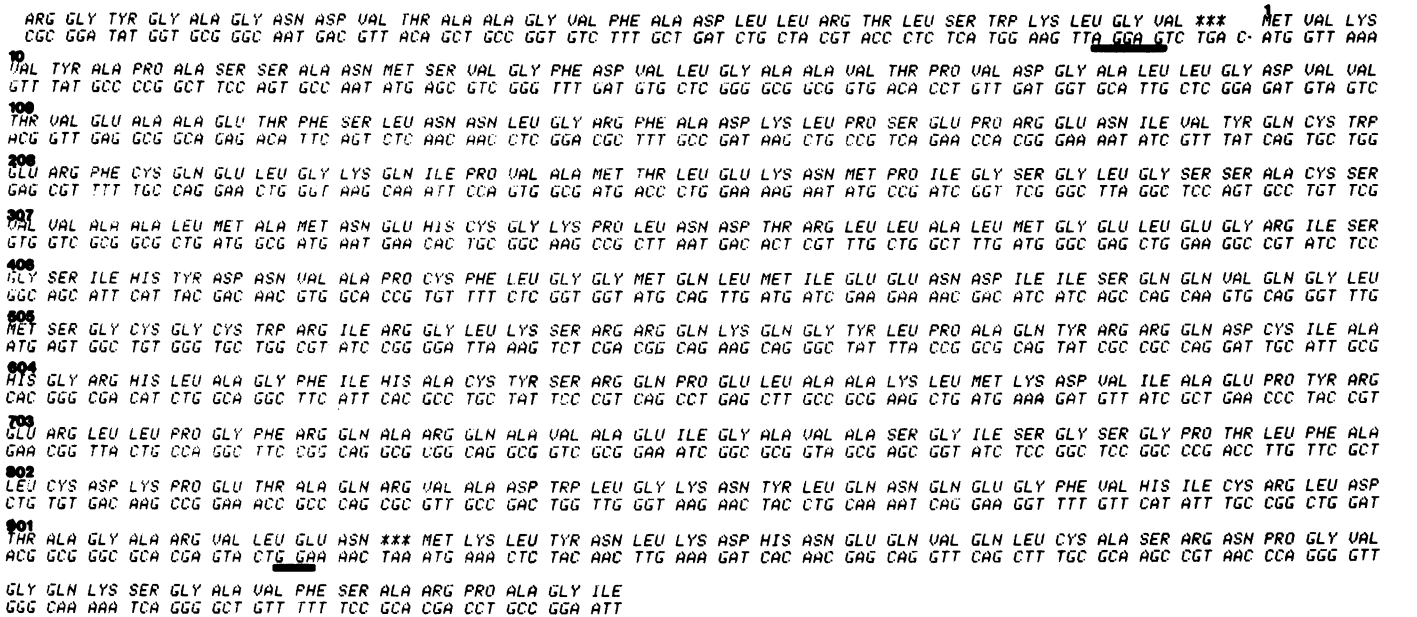

Figure 2. Nucleotide sequence of the thrB gene and the neighbouring sequences.

The nucleotide and aminoacid sequences of the encoded homoserine kinase as well as that of the C-terminal of aspartokinase I - homoserine dehydrogenase I and the presumed NH_2-terminal of threonine synthase are presented. The DNA sequence is numbered from the initiator codon ATG of thrB. The putative ribosome binding sites at the end of thrA and thrB are underlined.

rimidine) for Gly (15), GAA for Glu (16), CCG for Pro (17), GCG for Ala (18,19). A net preference is observed for CAG over CAA although the concentration of $tRNA^{Glu}_{CAG}$ exceeds only slightly that of $tRNA^{Glu}_{CAA}$ (20).

b) The thrAB and thrBC junctions.

As shown in figure 2, there is only one base pair between the opale nonsense codon at the end of thrA, and the initiation AUG codon of thrB. The DNA sequence reveals that a second nonsense codon UAA in phase with the UGA is found 6 base pairs further in the sequence. The determination of a unique carboxyl terminal sequence (21) for aspartokinase homoserine dehydrogenase I indicates that the UGA codon is very effective in the termination of translation. Despite the fact that UGA is usually considered as the most common termination signal (32), the presence of the ochre codon may be a security in the case of an opale suppressor in the cell.

The thrB gene is ending by UAA, the ochre nonsense codon and is immediately followed by an initiation codon and a 114 base pair sequence, in an open reading frame. No other open reading

Table 1. Codon usage in thrB.

	U	C	A	G
U	UUU Phe 5 UUC Phe 4 UUA Leu 4 UUG Leu 7	UCU Ser 1 UCC Ser 6 UCA Ser 1 UCG Ser 2	UAU Tyr 5 UAC Tyr 3 UAA Ochre 1 UAG Ambre 0	UGU Cys 4 UGC Cys 7 UGA Opal 0 UGG Trp 3
C	CUU Leu 2 CUC Leu 5 CUA Leu 0 CUG Leu 13	CCU Pro 2 CCC Pro 1 CCA Pro 3 CCG Pro 8	CAU His 3 CAC His 3 CAA Gln 3 CAG Gln 14	CGU Arg 6 CGC Arg 4 CGA Arg 3 CGG Arg 7
A	AUU Ile 5 AUC Ile 10 AUA Ile 0 AUG Met 11	ACU Thr 1 ACC Thr 3 ACA Thr 2 ACG Thr 2	AAU Asn 6 AAC Asn 6 AAA Lys 2 AAG Lys 9	AGU Ser 4 AGC Ser 4 AGA Arg 0 AGG Arg 0
G	GUU Val 8 GUC Val 4 GUA Val 3 GUG Val 6	GCU Ala 4 GCC Ala 8 GCA Ala 5 GCG Ala 16	GAU Asp 7 GAC Asp 5 GAA Glu 14 GAG Glu 5	GGU Gly 9 GGC Gly 14 GGA Gly 3 GGG Gly 4

frame after an ATG is present before the Eco RI site. There are no protein data available to determine if the deduced protein sequence is the sequence of the mature threonine synthase. However, the long open reading frame after an initiation codon preceded by a Shine and Dalgarno sequence (see Discussion) is a good indication that the ATG which follows the nonsense codon is the initiation codon for thrC. If the beginning of thrC does not lie in that region, the thrBC junction will be among the longest one found so far in E.coli operons.

DISCUSSION.

The threonine operon is composed of three structural genes thrA, thrB, thrC which are transcribed in that order (22). If we assume that the beginning of thrC is located just after the nonsense codon of thrB, we are in presence of an operon with very short intergenic sequences, just one base pair for the thrAB junction and none for the thrBC junction. The intercistronic sequences described so far in operons of E.coli are all very different in length ranging from 413 base pairs between rplA and rplJ (23), 65 base pairs (between lacY and lacA) and 54 base pairs (between lacZ and lacY) (24) to the extreme situation overlapping nonsense and initiation codons UGAUG in the tryptophan operon at the trpBA junction. These regions apparently do not share much similarity in their nucleotide sequence except for partial homology observed between the GalE-GalT junction of E.coli and the trpC-trpB junction of Salmonella typhimurium (26). However, in all the sequences known complementary sequences to the 16 S ribosomal rRNA are present before the beginning of the second gene. In the case of the threonine operon, such sequences are also found : AGGAG at the end of thrA, and GGA at the end of thrB. These sequences raise at least two questions : (i) are these ribosomal binding sites functional in vivo ? (ii) what is the fate of the ribosomes translating the first gene : do they continue to translate thrB or do they dissociate before initiation on thrB ? The partial polar effects of nonsense mutations in thrA on the expression of thrB and thrC are a good indication that the internal ribosomal binding site could function in vivo, at least when such mutations are present. Further experiments

A : Arg Val Ala Asp Ile Leu Glu Ser Asn Ala Arg

B : Arg Val Ala Asp Trp Leu Gly Lys Asn Tyr Leu

Gln Gly Gln

Gln Asn Gln

Figure 3. Sequence of the two peptides which have similar sequences in aspartokinase I homoserine dehydrogenase I (A) and homoserine kinase (B).

are necessary to answer the second question.

We were interested in comparing the thrB nucleotide sequence to that of thrA and see if those genes which belong to the same biosynthetic operon have derived from a common ancestor according to the hypothesis of Horowitz (27,28). Extensive computer analysis did not show any significant homology. The same analysis was performed on the gene products, the aspartokinase I-homoserine dehydrogenase I and the homoserine kinase, which have a common effector, the L-threonine. The only significant similarity found was between the aminoacids 19 to 33 in aspartokinase I homoserine dehydrogenase I and the aminoacids 276-289 in homoserine kinase, as shown in figure 3 where 8 aminoacids out of 14 are identical. The secondary structure of homoserine kinase predicted according to Garnier (29) shown in Figure 4 did not show any similarity with that of AKI-HDHI (12).

Different immunological approaches carried out on aspartokinase I homoserine dehydrogenase I and homoserine kinase led to divergent results on a common origin between the two proteins

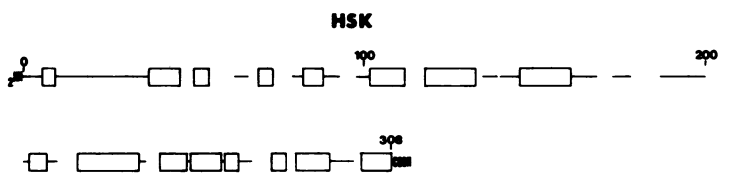

Figure 4. Predicted secondary structure of the homoserine kinase.

The boxes correspond to the possible α-helix structures, the lines to the possible extended regions of the protein. The coordinates above are the aminoacids from 1 to 820.

(30,31).

The homology presented here is so small that it leaves unsolved the question : have the two first genes a common origin ? The DNA sequence of the third gene of the operon, thrC coding for threonine synthase may help to clarify the origin of the three cistrons.

ACKNOWLEDGEMENTS.

We are grateful to R. Staden, F. Schaeffer, R. Garnier and J. Ninio for making available to us their computor programs and to P. Herbomel and B. Caudron for their help in their use. M.K. is extremely grateful to G. Winter and F. Sanger for the opportunity of learning the dideoxy-terminator sequencing technique. We thank G. Cohen, I. Saint Girons and B. Burr for valuable discussion and C. Maczuka for her patience in the preparation of the manuscript.

This work was supported by grants from the Centre National de la Recherche Scientifique (LA 270 and ATP : "Séquence des acides nucléiques informationnels"), the Institut National de la Santé et de la Recherche Médicale (ATP 77-82) and the Délégation Générale à la Recherche Scientifique et Technique. M.K. was supported by a short term E.M.B.O. fellowship.

REFERENCES.

1. Cossart, P., Katinka, M., Yaniv, M., Saint Girons,I. and Cohen, G.N. (1979) Molec.gen.Genet. 175, 39-44.
2. Maxam, A. and Gilbert, W. (1977) Proc.Natl.Acad.Sci. USA 74, 560-564.
3. Sanger, F., Nicklens, S. and Coulson, A. (1977) Proc.Natl. Acad.Sci. USA 74, 5463-5467.
4. Berkner, K.L. and Folk, W.R. (1977) J.Biol.Chem. 252, 3176-3184.
5. Schreier, P.H. and Cortese, R. (1979) J.Mol.Biol. 129, 169-172.
6. Rothstein, R.J., Lau, L.F., Bahl, C.P., Narang, S.A. and Wu, R. (1980) Methods in Enzymology 68, 98-109.
7. Sanger, F. and Coulson, N.A. (1978) FEBS Lett. 87, 107-110.
8. Staden, R. (1977) Nucl.Acids Res. 4, 4037-4051.
9. Staden, R. (1978) Nucl.Acids Res. 5, 1013-1015.
10. Staden, R. (1979) Nucl.Acids Res. 6, 2601-2610.
11. Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) J.Mol. Biol. 120, 97-120.

12. Katinka, M., Cossart, P., Sibilli, L., Saint Girons,I., Chalvignac, M.A., Lebras, G., Cohen, G.N. and Yaniv, M. (1980) Proc.Natl.Acad.Sci. USA, 77, 5730-5733.
13. Burr, B., Walker, J., Truffa-Bachi, P. and Cohen, G.N. (1976) Eur.J.Biochem. 62, 519-526.
14. Blank, H.U. and Söll, D. (1971) J.Biol.Chem. 246, 4947-4956.
15. Fleck, E.W. and Carbon, J. (1975) J.Bacteriol. 122, 492-501.
16. Ohashi, Z., Saneyoshi, M., Harada, F., Hara, H. and Nishimura, S. (1970) Biochem.Biophys.Res.Commun. 40, 866-872.
17. Söll, D., Cherayil, J.D. and Bock, R.M. (1967) J.Mol.Biol. 29, 97-112.
18. Williams, R.J., Nagel, W., Roe, B. and Dudock, B. (1974) Biochem.Biophys.Res.Commun. 60, 1215-1221.
19. Lund, E. and Dahlberg, J.E. (1977) Cell 111, 247-262.
20. Yaniv, M., Folk, W.R., Berg, P. and Soll, L. (1974) J.Mol.Biol. 86, 245-260.
21. Falcoz-Kelly, F., Janin, J., Saari, J.C., Veron, M., Truffa-Bachi, P. and Cohen, G.N. (1972) Eur.J.Biochem. 28, 507-519
22. Thèze, J. and Saint Girons, I. (1974) J.Bact. 118, 990-998.
23. Post, L., Strycharzg, D., Nomura, M., Lewis, H. and Dennis, P. (1979) Proc.Natl.Acad.Sci. 76, 1697-1701.
24. Büchel, D.E., Gronenborn, B. and Müller-Hill, B. (1980) Nature 283, 541-545.
25. Platt,T. and Yanofsky, C. (1975) Proc.Natl.Acad.Sci. USA 72, 2399-2403.
26. Selker, E. and Yanofsky, C. (1979) J.Mol.Biol. 130, 135-143.
27. Horowitz, N.H. (1945) Proc.Natl.Acad.Sci. USA 31, 153-157.
28. Horowitz, N.H. (1955) in "Evoling genes and proteins", Bryson, V. and Vogel, H.J. eds., Academic Press, pp. 15-23.
29. Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) J.Mol.Biol. 120, 97-120.
30. Truffa-Bachi, P., Guiso, N., Cohen, G.N., Thèze, J. and Burr, B. (1975) Proc.Natl.Acad.Sci. USA 72, 1268-1271.
31. Zakin, M.M., Garel, J.R., Dautry-Varsat, A., Cohen, G.N. and Boulot, G. (1978) Biochemistry 17, 4318-4323.
32. Steege, D.A. and Söll, D. in "Biological regulation and development", R.R. Goldberger, ed. (1980), Plenum Press, New York and London.