Volume 13 Number 24 1985 Nucleic Acids Research

Complete nucleotide sequence of the *E. coli* N-acetylneuraminate lyase

Yasuhiro Ohta, Kunihiko Watanabe[1] and Akira Kimura[1]

Kyoto Research Laboratories, Marukin Shoyu Co. Ltd., Uji and [1]Research Institute for Food Science, Kyoto University, Uji, Kyoto 611, Japan

Received 21 November 1985; Accepted 29 November 1985

Abstract

The nucleotide sequence of the cloned DNA, 1,243 bp in length coding for N-acetylneuraminate lyase (N-acetylneuraminate pyruvate lyase ; NPL) of Escherichia coli has been determined. Nucleotide sequence and amino acid analysis have assigned the open reading frame for NPL, starting with the ATG near its 5'terminus. The molecular weight calculated from the predicted amino acid sequence was 32,640 daltons, being in good agreement with that of a NPL subunit estimated by the SDS-PAGE method and amino acid composition. Several signal sequences conserved in the promoter regions of E. coli were found in the npl gene. They were the Shine-Dalgarno sequence, the Pribnow box and the sequence coserved in the "-35 region" and they were separated to each other with preferable spacing for an efficient transcription. Downstream from the termination codon, the inverted repeat sequence was present, followed by 4 successive T's.

Introduction

N-acetylneuraminate lyase (N-acetylneuraminate pyruvate lyase, N-acetylneuraminic acid aldolase, EC 4.1.3.3.) coverts N-acetylneuraminic acid to pyruvate and N-acetyl-D-mannosamine. It can be used for the determination of N-acetylneuraminate by coupling with either lactate dehydrogenase (1, 2) or pyruvate oxidase (3, 4). This was found to be distributed in a wide variety of bacteria belonging to species such as Escherichia, Pseudomonas, Aerobacter, Proteus, Micrococcus, Sarcina, Brevibacterium, Corynebacterium, Arthrobacter, Bacillus, Bacterium, Vibrio, and Clostridium (5, 6, 7). As reported previously (7), we have been studied on the production of N-acetylneuraminate lyase by E. coli. To attempt to get potent NPL-producer by gene manipulation techniques, we cloned a 1.2 kb HindIII-EcoRI fragment of E. coli chromosomal DNA that contained npl gene onto vector plasmid pBR322 and designated it pMK6.

The simple restriction map was also established (8,19). In this study, we determined the complete nucleotide sequence of this HindIII-EcoRI fragment of pMK6 hybrid plasmid carrying *npl* gene.

Materials and Methods

Bacterial Strain

Strain *E. coli* K-12 {C600(F⁻ *hsdR hsdM recA*⁺ *thr leu thi lacY supE tonA*)} was used.

DNA sequencing procedure

Plasmid pMK6 was prepared on a large scale from cleared lysate by banding in a CsCl gradient. pMK6 was digested by *EcoRI* in combination with *HindIII* restriction endonucleases and the resulting 1,243 bp fragment was separated from the vector by electrophoresis in a 1.0% low-melting-temperature agarose gel. The fragment redigested with *StuI*, *AluI*, *EcoRV*, *HaeIII* and *ScaI* was inserted into the corresponding cloning sites of M13 mp18 and M13 mp19 phage vectors and subcloned according to the supplier's specifications (RCC Amersham). The resulting recombinant phage DNA was sequenced by the "dideoxy sequencing method" of Sanger et al (9).

Purification of NPL and amino acid sequence determination

N-acetylneuraminate lyase was purified from *E. coli* C600 cells transformed with pMK6. The purification method was essentially the same as reported previously (10). Edman degradation were carried out for the amino-terminal sequence of the purified NPL on a Applied Biosystem Model 470A protein sequencer. The carboxy-terminal sequence was analyzed with carboxypeptidase A and B as described by Ambler(11).

Determination of amino acid composition

Proteins were hydrolyzed *in vacuo* with 6N HCl at 110°C for 22, 48, 72 hrs. The composition of amino acids was analyzed with Hitachi Model 835 Amino Acid Analyzer. Threonine and serine were corrected for 5 and 10 % destruction, respectively, during hydrolysis. Valine and isoleucine were taken the value of 72 hr hydrolysate. Half-cystine was determined as cysteic acid after performic acid oxidation (12). Tryptophan was determined spectrophotometrically (13).

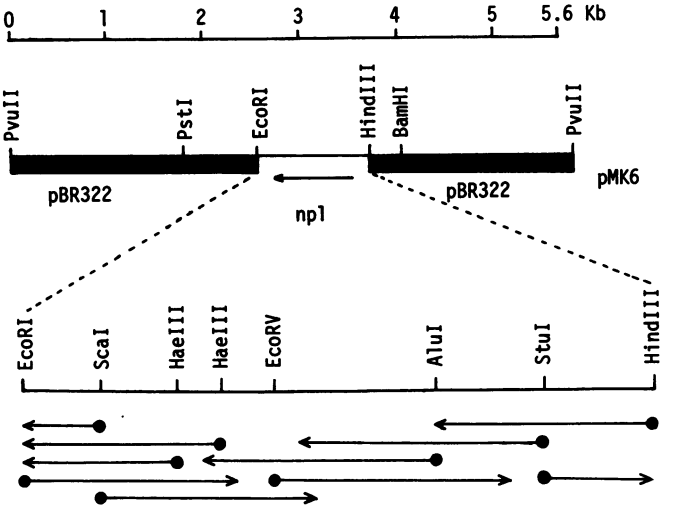

Figure 1. Physical map of the recombinant plasmid pMK6 and the strategy for the DNA sequencing of npl gene. The restriction sites of pMK6 DNA used for sequencing indicated at the map coordinated on the top column. At the bottom column in the expanding scale, the extent and direction of DNA sequencing are indicated by horizontal arrows.

Chemicals

Restriction endonucleases were obtained from Takara Shuzo Co. Ltd, Kyoto, Japan and Nippon Gene Co. Ltd, Toyama, Japan. Carboxypeptidase A and B were from Sigma Chemical Company, St.Louis, Mo. U.S.A. and Worthington Biochemicals, Co., Freehold, N.J., U.S.A., respectively. {α-^{32}P}dCTP was purchased from RCC Amersham. M13 cloning and sequencing kits were from Takara Shuzo Co. Ltd, Kyoto, Japan.

Results

Nucleotide sequence of the npl gene

The original hybrid plasmid pMK6 was 5.6 kb in length and contained the npl gene derived from E. coli K-12 chromosomal DNA fragment (1,243 bp) inserted at the EcoI-HindIII site of pBR322 (8). To determine the nucleotide sequence of the npl gene, we first digested the pMK6 by EcoRI in combination with HindIII to isolate the 1,243 bp DNA fragment containing the npl gene. The fragment was dissected further by restriction enzymes and

```
   1                                                                AAGCTTTCTGTATGGGGTGTTG

  23 CTTAATTGATCTGGTATAACAGGTATAAAGGTATATCGTTTATCAGACAAGCATCACTTCAGAGGTATTT
       -35 SQ                  Pribnow          starting of mRNA          SD
  93 ATG GCA ACG AAT TTA CGT GGC GTA ATG GCT GCA CTC CTG ACT CCT TTT GAC CAA
   1 Met Ala Thr Asn Leu Arg Gly Val Met Ala Ala Leu Leu Thr Pro Phe Asp Gln

 147 CAA CAA GCA CTG GAT AAA GCG AGT CTG CGT CGC CTG GTT CAG TTC AAT ATT CAG
  19 Gln Gln Ala Leu Asp Lys Ala Ser Leu Arg Arg Leu Val Gln Phe Asn Ile Gln

 201 CAG GGC ATC GAC GGT TTA TAC GTG GGT GGT TCG ACC GGC GAG GCC TTT GTA CAA
  37 Gln Gly Ile Asp Gly Leu Tyr Val Gly Gly Ser Thr Gly Glu Ala Phe Val Gln

 255 AGC CTT TCC GAG CGT GAA CAG GTA CTG GAA ATC GTC GCC GAA GAG GGC AAA GGT
  55 Ser Leu Ser Glu Arg Glu Gln Val Leu Glu Ile Val Ala Glu Glu Gly Lys Gly

 309 AAG ATT AAA CTC ATC GCC CAC GTC GGT TGC GTC ACG ACC GCC GAA AGC CAA CAA
  73 Lys Ile Lys Leu Ile Ala His Val Gly Cys Val Thr Thr Ala Glu Ser Gln Gln

 363 CTT GCG GCA TCG GCT AAA CGT TAT GGC TTC GAT GCC GTC TCC GCC GTC ACG CCG
  91 Leu Ala Ala Ser Ala Lys Arg Tyr Gly Phe Asp Ala Val Ser Ala Val Thr Pro

 417 TTC TAC TAT CCT TTC AGC TTT GAA GAA CAC TGC GAT CAC TAT CGG GCA ATT ATT
 109 Phe Tyr Tyr Pro Phe Ser Phe Glu Glu His Cys Asp His Tyr Arg Ala Ile Ile

 471 GAT TCG GCG GAT GGT TTG CCG ATG GTG GTG TAC AAC ATT CCA GCC CTG AGT GGG
 127 Asp Ser Ala Asp Gly Leu Pro Met Val Val Tyr Asn Ile Pro Ala Leu Ser Gly

 525 GTA AAA CTG ACC CTG GAT CAG ATC AAC ACA CTT GTT ACA TTG CCT GGC GTA GGT
 145 Val Lys Leu Thr Leu Asp Gln Ile Asn Thr Leu Val Thr Leu Pro Gly Val Gly

 579 GCG CTG AAA CAG ACC TCT GGC GAT CTC TAT CAG ATG GAG CAG ATC CGT CGT GAA
 163 Ala Leu Lys Gln Thr Ser Gly Asp Leu Tyr Gln Met Glu Gln Ile Arg Arg Glu

 633 CAT CCT GAT CTT GTG CTC TAT AAC GGT TAC GAC GAA ATC TTC GCC TCT GGT CTG
 181 His Pro Asp Leu Val Leu Tyr Asn Gly Tyr Asp Glu Ile Phe Ala Ser Gly Leu

 687 CTG GCG GGC GCT GAT GGT GGT ATC GGC AGT ACC TAC AAC ATC ATG GGC TGG CGC
 199 Leu Ala Gly Ala Asp Gly Gly Ile Gly Ser Thr Tyr Asn Ile Met Gly Trp Arg

 741 TAT CAG GGG ATC GTT AAG GCG CTG AAA GAA GGC GAT ATC CAG ACC GCG CAG AAA
 217 Tyr Gln Gly Ile Val Lys Ala Leu Lys Glu Gly Asp Ile Gln Thr Ala Gln Lys

 795 CTG CAA ACT GAA TGC AAT AAA GTC ATT GAT TTA CTG ATC AAA ACG GGC GTA TTC
 235 Leu Gln Thr Glu Cys Asn Lys Val Ile Asp Leu Leu Ile Lys Thr Gly Val Phe

 849 CGC GGC CTG AAA ACT GTC CTC CAT TAT ATG GAT GTC GTT TCT GTG CCG CTG TGC
 253 Arg Gly Leu Lys Thr Val Leu His Tyr Met Asp Val Val Ser Val Pro Leu Cys

 903 CGC AAA CCG TTT GGA CCG GTA GAT GAA AAA TAT CAG CCA GAA CTG AAG GCG CTG
 271 Arg Lys Pro Phe Gly Pro Val Asp Glu Lys Tyr Leu Pro Glu Leu Lys Ala Leu

 957 GCC CAG CAG TTG ATG CAA GAG CGC GGG TGA GTTGTTTCCCCTCGCTCGCCCTACCGGGTG
 289 Ala Gln Gln Leu Met Gln Glu Arg Gly End                          IR

1017 AGGGGAAATAAACGCATCTGTACCCTACAATTTTCATACCAAAGCGTGTGGGCATCGCCCACCGCGGGAG
                      IR      termination of mRNA
1087 ACTCACAATGAGTACTACAACCCAGAATATCCCGTGGTATCGCCATCTCAACCGTGCACAATGGCGCGCA

1157 TTTTCCGCTGCCTGGTTGGGATATCTGCTTGACGGTTTTGATTTCGTTTTAATCGCCCTGGTACTCACCG

1227 AAGTACAAGGTGAATTC
```

inserted into the corresponding cloning sites of the M13 vector. After transfection in E. coli JM109 strain, the recombinant phages were propagated and the resulting phage DNA was submitted to sequencing (Fig. 1).

The sequence of the 1,243 bp chromosomal DNA containing the npl gene is shown in Figure 2. Examination of the nucleotide sequence shows only one possible open reading frame, starting at position 93 and terminating at position 983, which was sufficiently long enough to make a polypeptide of about 33 kd in molecular weight. In order to ascertain the initiation site of the npl, we determined the amino-terminal sequence of purified NPL. It was Met-Ala-Thr-Asn-Leu-Arg-Gly-Val----------Gln (Table 1). This sequence completely coincided with that of the first 54 amino acids predicted from the DNA sequence. Also to ascertain another C-termination site of the npl, we determined the carboxy terminal sequence of NPL and verified to be -Arg-Gly (Table 2). The open reading region can code for a polypeptide of 297 amino acids. The amino acid sequence predicted from the NPL is shown in Figure 2. Based on the amino acid analysis of the purified enzyme NPL and molecular weight calculated from the length of DNA, the amino acid composition was estimated as follows : Asp+Asn 23, Thr 15, Ser 14, Glu+Gln 38, Pro 11, Gly 28, Ala 25, Val 24, Met 7, Ile 17, Leu 34, Tyr 13, Phe 10, Lys 16, His 5, Arg 12, Cys 4, and Trp 1. This result showed in good accordance with that predicted for the nucleotide sequence (Table 3). The molecular weight calculated from the predicted amino acid sequence is 32,640 daltons, which agrees well with

Figure 2. Complete nucleotide sequence of npl gene and the flanking regulatory unit sequences. The nucleotide sequence of the gene is indicated and nucleotide are numbered from the HindIII site. The amino acid sequence of npl predicted from the sequences is given below. Several regulatory sequences flanking the npl gene are indicated with underlines. They include the Shine-Dalgarno (SD) sequences, the Pribnow box and the conserved sequence (-35 SQ) located at about -10 and -35 nucleotides, respectively, upstream from the starting point of mRNA synthesis which was tentatively assigned to position 59 or 60 as judged from the topology of the regulatory signal sequence described above. Downstream from the termination codon of the npl gene at position 984-986 the inverted repeat sequences (IR) are presented at position 1002-1022 and 1035-1045, followed by 4 successive T's which is a preferable site for the termination of mRNA synthesis.

Table 1. Coincidence of the amino acid sequence at the NH_2-terminus of NPL with that predicted from nucleotide sequence.

NH_2-terminal residue number	Codon	Amino acid residue	NH_2-terminal residue number	Codon	Amino acid residue
1	ATG	Met	28	CGT	Arg
2	GCA	Ala	29	CGC	Arg
3	ACG	Thr	30	CTG	Leu
4	AAT	Asn	31	GTT	Val
5	TTA	Leu	32	CAG	Gln
6	CGT	Arg	33	TTC	Phe
7	GGC	Gly	34	AAT	Asn
8	GTA	Val	35	ATT	Ile
9	ATG	Met	36	CAG	Gln
10	GCT	Ala	37	CAG	Gln
11	GCA	Ala	38	GGC	Gly
12	CTC	Leu	39	ATC	Ile
13	CTG	Leu	40	GAC	Asp
14	ACT	Thr	41	GGT	Gly
15	CCT	Pro	42	TTA	Leu
16	TTT	Phe	43	TAC	Tyr
17	GAC	Asp	44	GTG	Val
18	CAA	Gln	45	GGT	Gly
19	CAA	Gln	46	GGT	Gly
20	CAA	Gln	47	TCG	Ser
21	GCA	Ala	48	ACC	Thr
22	CTG	Leu	49	GGC	Gly
23	GAT	Asp	50	GAG	Glu
24	AAA	Lys	51	GCC	Ala
25	GCG	Ala	52	TTT	Phe
26	AGT	Ser	53	GTA	Val
27	CTG	Leu	54	CAA	Gln

the molecular weight of the NPL subunit (33 kd) estimated by SDS-PAGE (10). Since an apparent molecular weight of the native NPL estimated by molecular sieve chromatography is about 98 kd, native NPL seems to be a trimeric enzyme composed of

Table 2. Carboxy-terminal sequence analysis by carboxypeptidases

	Released amino acid(mol/mol protein)			
	- Gln	- Glu	- Arg	- Gly
Carboxypeptidase A digestion (37°C, 3 hr)	0	0	0	1.0
Carboxypeptidase A and B mixed digestion (37°C, 3 hr)	trace	trace	0.7	0.9

Table 3. Amino acid composition of the NPL of E. coli.

Amino acid	Amino acid analysis		Predicted for nucleotide sequence
	Residue per molecule	Nearest integer	integer
Asp	23.3[a]	23[a]	16
Asn			7
Thr	14.7	15	15
Ser	14.2	14	14
Glu	38.1[b]	38[b]	17
Gln			21
Pro	10.8	11	11
Gly	28.1	28	28
Ala	24.9	25	25
Val	24.1	24	24
Met	6.7	7	7
Ile	17.2	17	17
Leu	33.8	34	34
Tyr	13.0	13	13
Phe	10.3	10	10
Lys	16.3	16	16
His	4.9	5	5
Arg	12.0	12	12
Cys	3.8	4	4
Trp	1.0	1	1
Total	(Mw=32,640)	297	297

a; Value of Asp+Asn, b; Value of Glu+Gln.

Table 4. Codon usage in the E. coli npl gene.

Phe	UUU	4	Ser	UCU	3	Tyr	UAU	8	Cys	UGU	0
	UUC	6		UCC	2		UAC	5		UGC	4
Leu	UUA	3		UCA	0	End	UAA	0	End	UGA	1
	UUG	3		UCG	3	End	UAG	0	Trp	UGG	0
Leu	CUU	4	Pro	CCU	4	His	CAU	2	Arg	CGU	6
	CUC	5		CCC	0		CAC	3		CGC	5
	CUA	0		CCA	2	Gln	CAA	8		CGA	0
	CUG	19		CCG	5		CAG	13		CGG	1
Ile	AUU	6	Thr	ACU	3	Asn	AAU	3	Ser	AGU	3
	AUC	11		ACC	6		AAC	4		AGC	3
	AUA	0		ACA	2	Lys	AAA	13	Arg	AGA	0
Met	AUG	7		ACG	4		AAG	3		AGG	0
Val	GUU	4	Ala	GCU	3	Asp	GAU	13	Gly	GGU	11
	GUC	8		GCC	9		GAC	3		GGC	13
	GUA	7		GCA	5	Glu	GAA	12		GGA	1
	GUG	5		GCG	8		GAG	5		GGG	3

three identical subunits (10). Codon usage for E. coli NPL derived from DNA sequence data is shown in Table 4.

Transcriptional signals

Prokaryotic consensus sequences for transcriptional initiation have well been documented (14-16). In the precise DNA sequence shown in Figure 2, we found two hexanucleotides TTAATT and TATAAA at position 24 and 46, respectively, preceding to the initiation codon ATG at position 93. The former sequence matches in three out of six positions to the consensus sequence TTGACA conserved in the "-35 region" upstream from the initiation site (14). The latter sequence also agrees in five out of six nucleotide with the Pribnow box TATAAT (14). Furthermore, these consensus sequenses are separated from each other by 16 bp, presumably being the more preferable spacing for an efficient transcription (15). Regarding a transcriptional termination signal, it has not been studied so extensively. The inverted repeat sequence can be located at position 1002-1022 and 1035-1045, which can form two stable hairpin loop structures. These inverted repeats are immediately followed by T-rich sequence. This sequence arrangement is often seen in the prokaryotic terminal of mRNA (14).

Discussion

We have established the complete nucleotide sequence of *E. coli* K-12 *npl* gene. The DNA sequence upstream from the 5' - terminus of the *npl* gene (position 20-60) is slightly A-T rich, where the promoter is expected to be located. The sequence of putative "-35 region" and Pribnow boxes of the *npl* gene are both very similar to the known consensus nucleotide sequences (14). In "-35 region", another Pribnow box-like sequence TATAAC is found at position 36 . However, the sequence is separated from the "-35 region" by 7 bp, it is too narrow from the typical spacing of 17 nucleotides (14). Importance of the critical spacing between the two consensus sequence is further strengthened by the DNA of considering the direct contact sites of RNA polymerase with the promoter region on three-dimensional model (16). From these points, we assigned the sequence TATAAA at position 46 to be the Pribnow box. However, further biochemical and genetic experiments will be required to confirm our tentative promoter sequence as well as the transcriptional start point. Furthermore, about 10 nucleotide upstream from the initiation codon ATG, a potential Shine-Dalgarno sequence-like GAGG (17)is located (Fig. 2). About 20 bp downstream from the translational termination codon TGA, there exist GC-rich inverted repeat sequences followed by a stretch of successive T's.

A molecular weight calculated from the 297 amino acid residues encoded from the *npl* gene is 32,640 daltons, and is in good accordance with that estimated from amino acid composition and SDS-PAGE (33 kd). As the NPL is composed of three identical subunits (10), the subunit molecular weight (32.6 kd) calculated from the DNA sequence data fits even better to this estimate than does the value (33 kd) by SDS-PAGE. Codon usage in the *npl* gene shows no strong bias from the common tendency (Table 4). We can estimate the frequency of optimal codon usage (Fop) to be 0.66, according to the method of Ikemura and Ozeki (18). Thus, *npl* gene seems to be only moderately expressed in *E. coli* cells.

Acknowledgement

We wish to express their sincere gratitude to Dr. K. Murata, Research Institute for Food Science, Kyoto University and Dr. Y. Uchida, Kyoto Research Laboratories, Marukin Shoyu Co. Ltd., for their discussion and suggestions throughout and to Assoc. Prof. M. Sakaguchi, Research Institute for Food Science, Kyoto University, for his encouragement. We are grateful to Assoc. Prof. R. Hayashi, for his cordinal instructions for analyzing amino acid composition and carboxy-terminal sequence, and Prof. M. Takanami, Institute for Chemical Research of Kyoto University, for his encouragement. We thank Dr. Y. Tsukada and Dr. T. Sugimori, for criticizing the manuscripts and their encouragement.

References

1. Brunetti, P., Swanson, A. and Roseman, S. (1963) in Method in Enzymology (Colowick, S. P. and Kaplan, N. O. eds) Vol.VI,pp.465-473, Academic Press, Inc., New York.
2. Taniuchi, K., Miyamoto, Y., Uchida, Y., Chifu, K., Mukai, M., Yamaguchi, N., Tsukada, Y., Sugimori, T., Doi, K. and Baba, S. (1979) Jpn. J. Clin. Chem. 7, 403-410.
3. Kosaka, A. and Nakane, K. (1979) Proceeding of the Symposium on Clinical Physiology and Pathology, Vol.19, Japan Society of Clinical Chemistry, pp.202-203.
4. Sugahara, K., Sugimoto, K., Nomura, O. and Usui, T.(1980) Clin. Chim. Acta 108, 493-498.
5. Heimer, R. and Meyer, K. (1956) Proc. Natl. Acad. Sci. U.S.A. 42, 728-734.
6. PopeBnae, E. A. and Drew, R. M. (1957) J. Biol. Chem. 228, 673-683.
7. Uchida, Y., Tsukada, Y. and Sugimori, T. (1985) Agric. Biol. Chem. 49, 181-187.
8. Appl. Microbiol. Biotech. in preparation.
9. Sanger, F., Nicklen, S. and Coulson, A. R. (1977) Proc. Natl. Acad. Sci. U.S.A. 74, 5463-5467.
10. Uchida, Y., Tsukada, Y. and Sugimori, T. (1984) J. Biochem., 96, 507-522.
11. Ambler, R. P. (1967) in Method in Enzymology (Lorand, L. ed) Vol.XI, pp.436-445, Academic Press, Inc., New York.
12. Moore, S. (1963) J. Biol. Chem. 238, 235-237.
13. Gaitonde, M. K. and Dovey, T. (1970) Biochem. J. 117, 907-911.
14. Rosenberg, M. and Court, D. (1979) Ann. Rev. Genet. 13, 319-353.
15. Hawley, D. K. and McClure, M. R. (1983) Nucl. Acids Res. 11, 2237-2255.
16. Siebenlist, U., Simpson, R. B. and Gilbert, W. (1980) Cell, 20, 269-281.
17. Shine, J. and Dalgarno, L. (1975) Proc. Natl. Acad. Sci. U.S.A. 72, 784-788.
18. Ikemura, T and Ozeki, H. (1983) Cold Spring Harbor Symp. Quant. Biol. 47, 1087-1097.
19. Sugimori, T. Tsukada, Y. Ohta, Y. and Kimura, A. (1984) Japanese Patent Application No.181250.