Cloning and structural characterization of the genes coding for adenosylcobalamin-dependent methylmalonyl-CoA mutase from *Propionibacterium shermanii*

E. Neil MARSH, Norman McKIE, Nigel K. DAVIS* and Peter F. LEADLAY†

Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1QW, U.K.

The structural genes coding for both subunits of adenosylcobalamin-dependent methylmalonyl-CoA mutase from the Gram-positive bacterium *Propionibacterium shermanii* have been cloned, with the use of synthetic oligonucleotides as primary hybridization probes. The genes are closely linked and are transcribed in the same direction. Nucleotide sequence analysis of 4.5 kb of DNA encompassing both genes allowed us to infer the complete amino acid sequence of the two subunits: the β -subunit is the product of the upstream gene, and consists of 638 amino acid residues (M_r 69465) and the α -subunit consists of 728 amino acid residues (M_r 80147). There is a very close structural homology between the two subunits, reflecting the probable duplication of a common ancestral gene. A sequence present only in the α -subunit is significantly homologous to a portion of the sequence of the methylmalonyl-CoA-binding subunit of transcarboxylase from *P. shermanii* [Samols, Thornton, Murtif, Kumar, Haase & Wood (1988) *J. Biol. Chem.* **263**, 6461–6464], and this homologous region may form part of the CoA ester-binding site in both enzymes.

INTRODUCTION

Methylmalonyl-CoA mutase (EC 5.4.99.2) is an adenosylcobalamin-dependent enzyme that catalyses the structural rearrangement of succinyl-CoA (3-carboxypropionyl-CoA) into (2*R*)-methylmalonyl-CoA as part of the propionic acid fermentation (Overath *et al.*, 1962; Kellermeyer *et al.*, 1964). Rearrangements promoted by adenosylcobalamin (coenzyme B-12) have been intensively studied (for reviews see Babior & Krouwer, 1979; Dolphin, 1982), but they remain amongst the most enigmatic of enzyme-catalysed reactions. The first step in the catalysis is thought to involve the homolytic cleavage of the cobalt–carbon bond in the 5'-deoxyadenosylcobalamin cofactor (Abeles, 1979; Finke & Hay, 1984; Wollowitz & Halpern, 1984) to generate free radical intermediates, but there is no agreement about the nature of the subsequent steps. The lack of detailed structural information for any of these proteins has frustrated real progress in this area.

Methylmalonyl-CoA mutase from *Propionibacterium shermanii* is one of the most accessible of these enzymes (for a review see Rétey, 1982). It is composed of two polypeptides: an α -subunit (M_r 79000) and a β -subunit (M_r 67000) (Zagalak *et al.*, 1974; Francalanci *et al.*, 1986; Wölfe *et al.*, 1986). As normally isolated, the mutase is a mixture of apoenzyme and cofactor-bound forms, which can be resolved by f.p.l.c. (Francalanci *et al.*, 1986). Stereochemical aspects of the enzyme-catalysed rearrangement have been examined in detail (Rétey, 1982; Wölfe *et al.*, 1986), and some progress has been made in the design of plausible chemical model reactions (Wollowitz & Halpern, 1984), but the part played by the protein is obscure. A cofactor-containing

form of the mutase from *P. shermanii* has been crystallized in a form suitable for high-resolution X-ray crystallography (Marsh *et al.*, 1988).

Here we report the cloning and DNA sequence analysis of the structural genes coding for both subunits of the mutase from *P. shermanii*. The inferred primary structure of the protein provides unequivocal confirmation of the subunit structure, and reveals significant similarity between the two non-identical subunits. It also reveals a highly significant local homology between a portion of the C-terminal region of the α -subunit and the 12 S subunit of transcarboxylase from *P. shermanii*, which has been shown to be the site of methylmalonyl-CoA binding in that enzyme.

MATERIALS AND METHODS

Purification and sequence analysis of proteolytic fragments of methylmalonyl-CoA mutase

Reduced carboxymethylated mutase from *P. shermanii* (Francalanci *et al.*, 1986) (12 mg, 80 nmol) was dissolved in 1.5 ml of 0.5% NH_4HCO_3 and treated with endoproteinase Arg-C (Boehringer, Mannheim, Germany) (0.12 mg/ml) for 7 h at 25 °C. A second, equal, amount of endoproteinase Arg-C was then added and reaction was allowed to continue for a further 16 h. The extent of digestion was monitored by h.p.l.c. on a TSK 3000 SWG (Toya Soda, obtained through Anachem, Luton, Beds., U.K.) gel-filtration column. The peptides were subjected to gel filtration on a column of Sephadex G-75 (superfine grade) (Pharmacia, Uppsala, Sweden) to remove undigested material. The peptides were separated by h.p.l.c. on a reverse-phase column (Lichrosorb RP8), eluted with increasing concentrations of acetonitrile in

* Present address: Department of Genetics, John Innes Institute and A.F.R.C. Institute of Plant Science Research, Colney Lane, Norwich NR4 7UH, U.K.

† To whom correspondence should be addressed.

These sequence data have been submitted to the EMBL/GenBank Data Libraries under the accession number X14965.

0.1% (v/v) trifluoroacetic acid. Peptides (5–10 nmol) were partially sequenced by published manual methods (Chang, 1983). Subsequently the *N*-terminal sequence of each intact mutase subunit was obtained after separation of the citraconylated subunits, by using ion-exchange chromatography on a Mono Q column (Pharmacia), and removal of the citraconyl groups by treatment with 5% (v/v) formic acid for 2 h at room temperature. This was done by automated sequence analysis with an Applied Biosystems (Foster City, CA, U.S.A.) 470A gas-phase sequencer, and phenylthiohydantoin derivatives were identified by reverse-phase h.p.l.c. with an Applied Biosystems on-line system.

Synthetic oligonucleotides

By using the manual 'paper-disc' modification (Frank *et al.*, 1983; Matthes *et al.*, 1984) of the phosphotriester method, it is possible to synthesize well over 100 individual 17-mers simultaneously in 2 days (Matthes *et al.*, 1984). For eight selected peptides derived from mutase by digestion with endoproteinase Arg-C, a total of 90 alternative synthetic oligonucleotide probes was constructed, with consideration being given to the biased codon usage observed previously for genes from *P. shermanii* (Murtif *et al.*, 1985; N. K. Davis, unpublished work). Of the 18 'best-guess' oligonucleotides based on the sequence determined for one such peptide, A13-14/8, one (M18), which hybridized strongly to Southern blots of *P. shermanii* genomic DNA under stringent conditions, is shown in Fig. 1. For use as hybridization probes, oligonucleotides were end-labelled with [γ - 32 P]ATP (5000 Ci/mmol; Amersham International, Amersham, Bucks., U.K.) and polynucleotide kinase (gift from Dr. T. Hunt) (Maniatis *et al.*, 1982, pp. 122–123) to a specific radioactivity of approx. 8×10^8 c.p.m./ μ g. Oligonucleotides for use in dideoxy sequencing were synthesized in a Biosearch Cyclone DNA synthesizer.

Construction and screening of *P. shermanii* libraries

Genomic DNA was purified from anaerobically grown *P. shermanii* (N.C.I.B. 9885) (Brown *et al.*, 1973). Cells (2 g wet wt.) were thoroughly ground to a powder under liquid N₂ with a mortar and pestle, and DNA was extracted according to published methods (Maniatis *et al.*, 1982). The genomic DNA was used to prepare a mini-library of *Eco*RI fragments (1.0–1.8 kb). DNA was digested to completion with *Eco*RI (Amersham International) and subjected to preparative electrophoresis on agarose gel. The appropriate slice was excised, and the DNA was recovered by binding to Whatman DE 81 paper (Dretzen *et al.*, 1981). The purified restriction fragments were inserted into the *Eco*RI site of plasmid pUC13 (Vieira & Messing, 1982) and used to transform *Escherichia coli* TG1 (Biggen *et al.*, 1983) by the method of Hanahan (1983). Positive colonies were identified from the library of *Eco*RI fragments by hybridization of oligonucleotide probes, end-labelled with 32 P, to colony replicas on filters of Whatman 541 filter paper prepared essentially as described by Maas (1983) except that colonies were pretreated with lysozyme. Hybridizations and washes were performed at approx. 7 °C below the calculated T_m (Suggs *et al.*, 1982) for each oligonucleotide. A *Kpn*I mini-library was constructed in an essentially similar manner except that *Kpn*I fragments (6.0–7.0 kb) were recovered by binding to NA 45 membranes (Schleicher and Schuell), inserted into the *Kpn*I site of

plasmic pUC18 (Vieira & Messing, 1982) and transformation of *E. coli* TG1 (Biggen *et al.*, 1983) was by the CaCl₂ method (Maniatis *et al.*, 1982). The *Kpn*I library was screened with the insert from plasmid pND4 (as above), previously 32 P-labelled by nick translation (Maniatis *et al.*, 1982). The initial plasmid isolate pND4 contains a 1.4 kb *Eco*RI fragment in the *Eco*RI site of plasmid pUC13. The plasmid isolate pMK1 contains a 6.6 kb *Kpn*I fragment in the *Kpn*I site of plasmid pUC18.

DNA sequencing and analysis

Specific restriction fragments were subcloned into plasmids M13 mp18 and M13 mp19 (Messing, 1983) or alternatively into the Bluescript plasmid vectors (Stratagene), and thence into single-stranded DNA according to the supplier's instructions (Bluescript Instruction Manual, Stratagene, San Diego, CA, U.S.A.). DNA was sequenced by the dideoxy chain-termination method (Sanger *et al.*, 1977) with either universal primer or specific oligonucleotide primers. Plasmid DNA for double-stranded sequencing was banded twice on CsCl gradients before use, and sequencing was carried out with modified T7 DNA polymerase (Sequenase) essentially according to the manufacturer's instructions (Zhang *et al.*, 1988) but adapted for high-G + C-content DNA (N. Dhillon & P. F. Leadlay, unpublished work). Sequence data were compiled and analysed by using the computer programs of Staden (1982, 1984). Secondary-structure prediction was calculated by using the method of Chou & Fasman (1978). Hydropathy plots of the mutase sequences were done by the method of Kyte & Doolittle (1982). Sequence comparisons were made using either the DIAGON program of Staden (1984) or the ALIGN program available from the Protein Identification Resource of the National Biomedical Research Foundation (Washington, DC, U.S.A.) and based on the algorithm of Needleman & Wunsch (1970).

RESULTS AND DISCUSSION

Cloning of the genes coding for the two subunits of methylmalonyl-CoA mutase

Endoproteinase Arg-C was used to digest reduced carboxymethylated methylmalonyl-CoA mutase and obtain relatively large peptide fragments. Manual sequencing of purified peptides identified eight relatively short sequences that could be used for making oligonucleotide probes. Advantage was taken, in initial screening, of the rapid 'paper-disc' method of manual oligonucleotide synthesis: 90 individual oligonucleotides were synthesized in a single experiment, based on the sequences of the eight peptides and with allowance for the biased codon usage found in these bacteria (Murtif *et al.*, 1985; N. K. Davis, unpublished work). This approach circumvented the problems of using complex mixtures of 'best-guess' oligonucleotides, and of relying on a single manually determined peptide sequence for probe design.

Genomic DNA from *P. shermanii* was cut with *Eco*RI and subjected to electrophoresis in wide tracks on agarose gels. The gels were dried on Whatman 3MM paper and cut into strips, which were hybridized separately with each oligonucleotide, previously end-labelled with 32 P. Oligonucleotide M18 (for the sequence see Fig. 1) hybridized strongly to a 1.4 kb *Eco*RI fragment of *P. shermanii* DNA, and six other oligonucleotide probes

also bound, although less convincingly. These seven probes were used in hybridization experiments with genomic DNA cut with a selection of restriction enzymes. Again, the most noticeably specific hybridization was with oligonucleotide M18, and this probe was chosen for hybridization to a *P. shermanii* library consisting of 1.0–1.8 kb *EcoRI* fragments in the plasmid pUC13. Several positive transformants were identified, and one, designated pND4, was selected for further analysis.

Limited DNA sequence analysis of the insert in pND4 by the random M13 dideoxy chain termination method (Biggen *et al.*, 1983) showed the presence of a DNA sequence exactly complementary to oligonucleotide M18 and entirely consistent with the amino acid sequence of peptide A13-14/8 used in the design of this probe (Fig. 1). A DNA sequence was also present corresponding to the *N*-terminus of the larger subunit of the mutase, and 3' to that matching M18. This suggested that the insert in pND4 contained the end of one gene and the start of the other. The insert in pND4 was labelled with ^{32}P by nick translation (Maniatis *et al.*, 1982) and used as a hybridization probe for a *P. shermanii* library consisting of 6.0–7.0 kb fragments in the plasmid pUC18. Several positive transformants were identified, and one, pMK1, was selected for further analysis. The restriction map of the insert in pMK1 is shown in Fig. 2, from which it is clear that the insert is large enough to accommodate the two closely linked genes for methylmalonyl-CoA mutase.

DNA sequencing strategy

Much of the DNA sequencing was carried out on

<i>N</i> -Terminal	Ser	Leu/Ile	Glu	Thr	Lys	Pro ...
mRNA 5'	UCG	UAG	CUC	UGG	UUC	GG ...
Oligo M18 5'	TCG	TAG	CTC	TGG	TTC	GG 3'

Fig. 1. *N*-Terminal sequence of a peptide (A13-14/8) isolated from an endoproteinase Arg-C digest of reduced carboxymethylated methylmalonyl-CoA mutase and used to design synthetic oligonucleotide probes encoding that sequence

One of these probes (M18), whose sequence is also shown, hybridized strongly to restriction digests of total DNA from *P. shermanii*.

EcoRI restriction fragments of the insert in MK1: the 2.2 kb and 1.0 kb fragments were subcloned in both orientations into the Bluescript plasmid SK⁺, from which single-stranded template was obtained by using helper bacteriophage. The region upstream of the mutase genes was sequenced after excision of an 800 bp *XhoI* fragment, and the 1.4 kb *EcoRI* insert also present in pND4 was subcloned as two *SalI*–*EcoRI* fragments into M13 mp18 and M13 mp19. The final piece of sequence encoding the C-terminus of the α -subunit was determined by double-stranded sequencing of a plasmid derived from pMK1 by deletion of the 1.0 kb, 1.4 kb and 2.2 kb *EcoRI* fragments. The overall strategy is summarized in Fig. 2. The restriction sites used to generate fragments for sequence analysis were confirmed by double-stranded sequencing of the parent plasmid, pMK1. In all, some 4.5 kb of sequence was obtained, spanning the region from 300 nucleotide residues upstream of the first gene to 60 bp 3' of the stop codon for the second gene. The sequence was fully determined on both strands throughout the coding region. The nucleotide sequence is shown together with the deduced amino acid sequence in Fig. 3.

Analysis of the nucleotide sequence

The region upstream of the initiation codon of the first gene shows no recognizable sequence similarity to consensus promoter sequences in other prokaryotes. However, an *E. coli*-like potential ribosome-binding site can be readily discerned before the start of each gene, and when the genes are placed on a plasmid behind a strong *E. coli* promoter both mutase genes are expressed *in vivo* in *E. coli* (N. McKie, unpublished work). The two genes lie very close together with the last G of the C-terminal lysine of the *mutA* gene being the first base of the GTG initiation codon of the *mutB* gene, which suggests that the genes are probably transcribed as an operon. A region of inverted symmetry, which is a potential transcription termination structure, is present 20 bp downstream of the termination codon of the second gene. The first gene (*mutA*) comprises 638 codons (1914 nucleotide residues) and encodes a protein of M_r 69465. This corresponds to the β -subunit of methylmalonyl-CoA mutase, which has an M_r of 67000 as estimated by SDS/polyacrylamide-gel electrophoresis. The second gene (*mutB*) is 728 codons (2184 nucleotide residues) in length and encodes a protein of M_r 80147. This

Fig. 2. Restriction map of the 6.6 kb *KpnI* fragment encoding the genes for methylmalonyl-CoA mutase (open boxes), showing the restriction sites used in the subcloning of fragments for DNA sequence determination

The arrows below indicate representative sequencing gels primed either from universal primer (↔) or from specifically-synthesized oligonucleotide primers (→).

GCAGCAGCAACATCTGTTGACGCCCTCGCGTTTCCAGATGTCCGATCCGGCCCCGGGAGGGCCGTGTTTGGCATTTCGGGGACGATTTCAGATGCCGAGGAGCACCTGTGAATCACATTGTC
 10 20 30 40 50 60 70 80 90 100 110 120

GGCAACGTGGCCCGCAAGCACGCCCCACGCAGCGGAGTTGACGTTTCGCTCCCTGCCACTGCGGCGCAGCCCCGGGACCGTGGGTCCCGACCCCGGACGGCTGATCGCCGCC
 130 140 150 160 170 180 190 200 210 220 230 240

ACGAGGACCCGATCGCGCTGTTGGTGGTGCCGTGCCCGCATGGCATGCCCGGCATCGCATCGCCGGGCAATCAGTTACCGGCACCCCTCGAGGTGCCTTGACCCCTTTCAGGAGCAGGAGT
 250 260 270 280 290 300 310 320 330 340 350 360

mutA
 M S S T D O G T N P A D T D D L T P T T L S L A G D F P K A T E E Q W E R E V
 CGCATGAGCAGCAGGATCAGGGGACCAACCCCGCGACACTGACGACCTCACTCCACCCACACTCAGCCTGGCCGGGATTTCCTCAAGGCCACTGAGGAGCAGTGGGAGCGCGAAGTT
 370 380 390 400 410 420 430 440 450 460 470 480

E K V L N R G R P P E K Q L T F A E C L K R L T V H T V D G I D I V P M Y R P K
 GAGAAGTACTCAACCGTGGTTCGCCAGGAGAGCAGTTGACCTTCGCGGAGTGTCTGAAGCGCCTGACGGTTACACCGCTCGATGCCATCGTCCGATGTACCGTCCGAAG
 490 500 510 520 530 540 550 560 570 580 590 600

D A P K K L G Y P G V A P F T R G T T V R N G D M D A W D V R A L H E D P D E K
 GACGCCCCGAAGAAGTGGGTATCCCGCGTGCACCTTTACCCCGCGCACCGGTGCGCAACCGCGACATGGATGCCTGGGACGTGCGCGCCTGCACGAGGATCCCGACGAGAAG
 610 620 630 640 650 660 670 680 690 700 710 720

F T R K A I L E G L E R G V T S L L L R V D P D A I A P E H L D E V L S D V L L
 TTCACCCGCAAGGCGATCCTCGAAGGCTGGAGCGTGGCGTCACCTCCCTGTTGTGCGCGTTGATCCCGACGCGATCGCACCCGAGCACCTCGACGAGGTCTCTCCGAGCTCTGCTG
 730 740 750 760 770 780 790 800 810 820 830 840

E M T K V E V F S R Y D Q G A A A E A L V S V Y E R S D K P A K D L A L N L G L
 GAAATGACCAAGTGGAGGTCTTCAGCCGTACGACAGGGTGC CGCGCGGAGGCCCTGGTGAGCGTCTACGAGCGCTCCGACAAGCCGCGAAGGACCTGGCCCTCAACCTGGGCGT
 850 860 870 880 890 900 910 920 930 940 950 960

D P I A F A A L Q G T E P D L T V L G D W V R R L A K F S P D S R A V T I D A N
 GATCCCATCGCGTTTCGACGCCCTGCAGGGCACCAGCGGATCTGACCGTGCTCGGTGACTGGTGCGCGCCCTGGCGAAGTTCTCGCCGACTCGCGCGCGCTCAGCATCGACGCGAAC
 970 980 990 1000 1010 1020 1030 1040 1050 1060 1070 1080

I Y H N A G A G D V A E L A W A L A T G A E Y V R A L V E Q G F T A T E A F D T
 ATCTACCACAACCGCGGTGCCCGCAGCTGGCAGAGCTCGCCTGGGCACTGGCCACCGCGCGGAGTACGTGCGCGCCCTGGTTCGAGCAGGGCTTACCGCCACCGAGGCTTTCGACAG
 1090 1100 1110 1120 1130 1140 1150 1160 1170 1180 1190 1200

I N F R V T A T H D Q F L T I A R L R A L R E A W A R I G E V F G V D E D K R G
 ATCAACTTCGGTTCACCGCCACCCACGACAGTCTCTCAGATCGCCCGCTTTCGCGCCCTGCGCGAGGATGGGCCCGCATCGCGAGGTCTTCGGCGTGGACGAGGACAAGCGCGG
 1210 1220 1230 1240 1250 1260 1270 1280 1290 1300 1310 1320

A R Q N A I T S W R D V T R E D P Y V N I L R G S I A T F S A S V G G A E S I T
 GCCCGCAGAATCGCATCACCAGCTGGCGTGACGTACGCGCGAAGACCCCTATGTCAACATCCTTCGCGGTTTCGATTGCCACCTTCTCCGCTCCGTGGTGGGCGGAGTTCGATCAG
 1330 1340 1350 1360 1370 1380 1390 1400 1410 1420 1430 1440

T L P F T Q A L G L P E D D F P L R I A R N T G I V L A E E V N I G R V N D P A
 AGCGTCCCTTACCCAGGCCCTCGCGCTGCGGAGGACGACTTCCCGCTGCGCATCGCGCGCAACACGGGATCGTGTCTCGCCGAGAGGTGAACATCGGCGCGCTCAACGACCGCGCC
 1450 1460 1470 1480 1490 1500 1510 1520 1530 1540 1550 1560

G G S Y Y V E S L T R S L A D A A W K E F Q E V E K L G G M S K A V M T E H V T
 GGTGGCTCCTTACTACGTGCGTCAACCGCAGCTGGCGCAGCGCGCTGGAAGGAATTCCAGGAGTTCGAGAAGCTCGGTGGCATGTGCGAAGGCCGTATGACCGAGCAGCTCACC
 1570 1580 1590 1600 1610 1620 1630 1640 1650 1660 1670 1680

K V L D A C N A E R A K R L A N R K Q P I T A V S E F P M I G A R S I E T K P F
 AAGTGTCTGACGCTGCAATGCCGAGCGCGCAAGCGCTGGCCAAACCGCAGCGAGCTACCGCGGTGACGAGTTCGCGATGATCGGGGCCGCGAGCATCGAGACCAAGCGGTT
 1690 1700 1710 1720 1730 1740 1750 1760 1770 1780 1790 1800

P A A P A R K G L A W H R D S E V F E Q L M D R S T S V S E R P K V F L A C L G
 CCCCGCGCTCCGCGCGCAAGGCCCTGGCCTGGCATCGCGACTCCGAGGTGTTTCGAGCAGCTGATGGATCGCTCCACCGAGCTCTCCGAGCGCCCCAAGGTGTCTTGGCTGCTTGGGC
 1810 1820 1830 1840 1850 1860 1870 1880 1890 1900 1910 1920

T R R D F G G R E G F S S P V W H I A G I D T P Q V E G G T T A E I V E A F K K
 ACCCGTCCGACTTCCGTTGGCCGCGAGGCTTCTCGAGCCCGGTGGGCACATCGCGCGCATCGACCCCCAGGTTCGAAGCGCGCACCCCGCCGAGATCGTCCGAGGATTCAGGAAG
 1930 1940 1950 1960 1970 1980 1990 2000 2010 2020 2030 2040

S G A Q V A D L C S S A K V Y A Q Q G L E V A K A L K A A G A K A L Y L S G A F
 TCGGCGCCCGAGTGGCGGACCTTCTGCTCGTCCGCAAGGTCTACGCGCAGCAGGACTTGAAGTCCGCAAGGCACTCAAGGCGCGCGCGCAAGGCCCTGTACCTGTGCGGCGCCTTC
 2050 2060 2070 2080 2090 2100 2110 2120 2130 2140 2150 2160

K E F G D D A A E A E K L I D G R L F M G M D V V D T L S S T L D I L G V A K * *mutB*
 AAGGAGTTCGGTGTGACGCGCGGAGGCGGAGAAGCTGATCGACGAGCGCTGTTTATGGGCATGGATGTGCTCGACACCTGTCCTCCACCTTGATATCTTGGAGTTCGCGAAGTGA
 2170 2180 2190 2200 2210 2220 2230 2240 2250 2260 2270 2280

T L P R F D S V D L G N A P V P A D A A R R F E E L A A K A G T G E A W E T A E
 GCACTCTGCCCCGTTTGTGATTGACCTCGGCAATGCCCGGTTCTGCTGATGCGCGCAGCGCTTCGAGGAAGTGGCGCCAAAGGCCGCGACCGGAGGCGTGGGAGACGCGCG
 2290 2300 2310 2320 2330 2340 2350 2360 2370 2380 2390 2400

Q I P V G T L F N E D V Y K D M D W L D T Y A G I P P F V H G P Y A T M Y A F R
 AGCAGATTCCGGTGGCACCTGTTCAACGAAGAGCTTACAAGGACATGGACTGGCTGGACACCTACGAGGTATCCCGCGTTCGTCACGCGCGCTGCAACCATGTACGCGTTC
 2410 2420 2430 2440 2450 2460 2470 2480 2490 2500 2510 2520

Fig. 3. DNA sequence of the genes encoding methylmalonyl-CoA mutase

Vol. 260

Table 1. Codon usage in *P. shermanii* *mutA* and *mutB* genes

The Table gives each codon with the corresponding amino acid in parentheses. Codon usage is expressed as percentage of all the codons for a particular amino acid calculated from both genes.

Codon	Usage overall (%)	Codon	Usage overall (%)	Codon	Usage overall (%)	Codon	Usage overall (%)
UUU (F)	4.0	UCU (S)	1.4	UAU (Y)	18.2	UGU (C)	16.7
UUC (F)	96.0	UCC (S)	31.0	UAC (Y)	81.8	UGC (C)	83.3
UUA (L)	0.0	UCA (S)	1.4	UAA (-)	0.0	UGA (-)	50.0
UUG (L)	3.5	UCG (S)	39.4	UAG (-)	50.0	UGG (W)	100.0
CUU (L)	4.3	CCU (P)	7.4	CAU (H)	20.0	CGU (R)	25.6
CUC (L)	30.4	CCC (P)	32.4	CAC (H)	80.0	CGC (R)	70.9
CUA (L)	0.0	CCA (P)	1.4	CAA (Q)	0.0	CGA (R)	2.3
CUG (L)	61.8	CCG (P)	58.8	CAG (Q)	100.0	CGG (R)	1.2
AUU (I)	12.3	ACU (T)	4.6	AAU (N)	19.4	AGU (S)	2.8
AUC (I)	87.7	ACC (T)	67.9	AAC (N)	80.5	AGC (S)	23.9
AUA (I)	0.0	ACA (T)	1.1	AAA (K)	1.5	AGA (R)	0.0
AUG (M)	100.0	ACG (T)	26.4	AAG (K)	98.5	AGG (R)	0.0
GUU (V)	14.7	GCU (A)	2.5	GAU (D)	22.1	GGU (G)	20.6
GUC (V)	45.6	GCC (A)	74.2	GAC (D)	77.9	GGC (G)	66.6
GUA (V)	0.9	GCA (A)	10.4	GAA (E)	16.0	GGA (G)	5.9
GUG (V)	38.8	CGC (A)	12.9	GAG (E)	84.0	GGG (G)	6.9

corresponds to the α -subunit of methylmalonyl-CoA mutase, which has an M_r of 79000 as estimated by SDS/polyacrylamide-gel electrophoresis. Again the value estimated by electrophoresis is slightly lower but agrees within the limits of experimental error. The identity of the gene products was confirmed by *N*-terminal protein sequencing of the separated α - and β -subunits as described in the Materials and methods section. The initiation codon for the first gene is ATG (methionine) and for the second gene it is GTG (normally a codon for valine). When the separated subunits were subjected to automated sequence analysis, no traces of methionine were found in the first round of degradation, suggesting that for both subunits methionine is post-translationally cleaved to produce the mature protein, as might be predicted from the criteria of Flinta *et al.* (1986).

Codon usage

Table 1 shows the codon usage as calculated from the nucleotide sequence of both genes. The preference for each codon is expressed as a percentage of all the codons for that amino acid. The results indicate that codon usage is non-random and, as expected for an organism with a high G+C content (64%), G or C is highly favoured in the third position where this is variable, as was also found for the gene encoding the 1.3 S biotin carboxy-carrier subunit of transcarboxylase from *P. shermanii* (Murtif *et al.*, 1985). However, codons that have been found to be strongly modulating in *E. coli* (Grosjean & Fiers, 1982) (those that correspond to rare or weakly binding tRNAs) are avoided, even where they would be favoured by DNA with a high G+C content, for example CGG for arginine and GGG for glycine. This is consistent with the genes being efficiently expressed in *P. shermanii*.

Sequence homologies

Comparison of the peptide sequences of the mutase genes by using the DIAGON program of Staden (1984)

Fig. 4. DIAGON plot showing the homology between the subunits of methylmalonyl-CoA mutase

The α -subunit is plotted horizontally and the β -subunit is plotted vertically. The window length was 51 amino acid residues and the score, calculated from the mutation data matrix of Schwartz & Dayhoff (1979) was set to exceed 565.

(Fig. 4) revealed the two polypeptides to be highly homologous, particularly towards the middle of the sequences, where about 80% of the amino acid residues are identical or substituted by similar residues. To assess more closely the homology between the two sequences, they were aligned by means of the protein alignment program ALIGN. The scoring matrix used was the mutation data matrix of Schwartz & Dayhoff (1979) with a bias of +5, and the penalty assigned to a padding

```

---MSSTDQGTNPADTDDLTPPTTSLSLAGDFPKATEFAEACLKRLVHTVDGIDIVPMYRPKADAPKKLGYPGVA
      |           |           |           |           |           |           |           |
VSTLPRFDSVDLGNAPVPADAARRFEELAAGAGTWGLDITYAGIPPFVHGYPATMYAFRPWTIRQYAGFSTAP

LHEDPDEKFTFRKAILEGLERGVTSLLRVDPDAIDAQGAAAEALVSVYERSDKPAKDLALNLGLDPIAFAS
      |           |           |           |           |           |           |           |
THRGYDSDNPRVAGDVGMAVDAISYDM-RELFVAVTAAEEQGVKPEQLAGTIQNDILKEFMVRNTYIYFPW

RAVTIDANIYHNAGAGDVAEALAWALATGAEYVRALFLTIARLRLREAWARIGEVFGVDEDKR-GARQNAITA
      |           |           |           |           |           |           |           |
NSISISGYHMQEAGATADIEMAYTLADGVDYIRAGFMEVAKLRAARMLWAKLVHQFGPKNPKSMSLRTHSQA

SVGGAESITTLPTFTQALGLPEDDFPLRIARNTGIVRSLADAANKEFQEVEKLGGMKSAVMTEHVTKVLADAG
      |           |           |           |           |           |           |           |
TQGHTQSLHTNSLDEAIALPTD-FSARIARNTQLFWDLARKAWGHIQEVEKVGGMKAIEKGIPKMRIEEAH

ARSIETKPFPAAPARKG-----LAWHRDSEVFEACLGRTRDFGGREG-----FSSPVWHIAGIG
      |           |           |           |           |           |           |           |
EPPLDVLKVDNSTVLAEQAKLVKLRAERDPEKVKLCIDAGRAMATVGEMSDALEKVFGRYTAQIRTISGVE

AQ-----VADLCSSAKVYAQQGLEVAKALKAAGAKRLFMG---MDVVDTLSSDILDILGVAK]
      |           |           |           |           |           |           |           |
GRRPRILLAKMGQDGHDRGQKVIAIYADLGFDDVSSLAGGHLTLVPLARKELDKLGRPDLILITVGGVIPS

VKKLRLASLDA

```

The ALIGN program was used as described in the text. By using a matrix bias of +5 and a gap penalty of -20, this alignment gave a score 16.7 standard deviations above the average score for the alignment of 100 random permutations of the two sequences. Pads are shown as - and amino acid residues identical in both sequences are indicated by vertical bars.

	563	573	583
Mutase α -subunit	A Q I R T I S G - - - - V Y S K E V K N T P E V E E E A R - E		
	+ *	+ + * * + *	+ + * *
Transcarboxylase	I A V M G A E G A A N V I F R K E I K A A D D P D A M R A E		
12 S subunit	431	441	451
	593	603	613
Mutase α -subunit	L V E E F E Q A E G R R P R I L L A K M G Q D G H		
	+ * * + + +	* *	* +
Transcarboxylase	K I E E Y Q N G S T R R T W R - - A R - G Q V - -		
12 S subunit	461	471	481

Pads are shown as –, amino acid residues identical in both sequences are indicated by ★, and conservative changes by +. The numbering is from A-563 of the mutase α -subunit and I-431 of the transcarboxylase 12 S subunit, and includes pads.

No sequences have previously been reported for a coenzyme B-12-dependent enzyme from any source, although the sequence of vitamin B-12-binding intrinsic factor has been determined (Dieckgraefe *et al.*, 1988). It was therefore not unduly surprising when screening of the proteins in either the PIR Protein Identification Resource Protein Sequence Database [(1986) National Biomedical Research Foundation Washington, DC,

U.S.A., release 10] or the SWISS-PROT Protein Sequence Data Bank [(1987) Departement de Biochimie Medicale, Centre Medical Universitaire, Geneva, Switzerland, release 5] did not reveal any striking sequence similarities to the mutase polypeptides. However, comparison of the mutase sequences with those of the constituent polypeptides of transcarboxylase from *P. shermanii* (Samols *et al.*, 1988), by using the ALIGN program, has revealed a significant local homology between these two enzymes. The homology is specifically between the 12 S subunit of transcarboxylase, which is known to contain the methylmalonyl-CoA-binding site (Wood & Zwolinski, 1976), and a portion of the C-terminal region of the α -subunit of mutase, which has no counterpart in the shorter β -subunit. The alignment score, for the 50-residue segments compared in Fig. 6, exceeded the average random score by 5.0 standard

deviations, indicating a genuine homology. The predicted secondary structure (Chou & Fasman, 1978) and hydrophathy plots (Kyte & Doolittle, 1982) for these two stretches of polypeptide are also very similar (results not shown). The majority of the identical or highly-conserved residues are charged, and we propose that this region may form part of the methylmalonyl-CoA-binding site in each of the two enzymes.

The sequences of other enzymes with acyl-CoA ester substrates did not show very convincing similarity to the sequences in Fig. 6, and this lack of a general sequence 'fingerprint' for acyl-CoA ester-binding sites has been previously noted (Buck *et al.*, 1985). Examination of the high-resolution X-ray-crystallographic structure for citrate synthase from chicken (Remington *et al.*, 1982; Wiegand *et al.*, 1984) shows that CoA binds to the enzyme in a compact form, between the two domains within each subunit. The adenine ring makes both hydrogen-bonding and hydrophobic contacts with the pantotheinyl chain, and eight of the 20 hydrogen bonds are to solvent water, rather than to amino acid residues that might have been conserved in this class of enzymes. An important contact is also made between the 3'-phospho group, and Arg-164 in the other subunit. Such considerations show the importance of obtaining three-dimensional structural information for the mutase, rather than attempting to extrapolate from the citrate synthase structure. Meanwhile, methylmalonyl-CoA mutase has now been cloned from human liver (Ledley *et al.*, 1988), and it will be of obvious interest to compare the sequence of that enzyme with the sequence determined here.

Note added in proof (received 7 April 1989)

The determination of the nucleotide sequences for the methylmalonyl-CoA mutases from both human and mouse liver has recently been completed (F. D. Ledley, personal communication).

We thank Dr. Brian Dunbar (University of Aberdeen, Aberdeen, U.K.) and Dr. John Findlay (University of Leeds, Leeds, U.K.) for the automated determination of protein sequences, and Margaret Frankland of the Wellcome Joint Pathology/Biochemistry Oligonucleotide Synthesis Facility (Department of Biochemistry, University of Cambridge) for oligonucleotide primers used in DNA sequencing. This research was supported by Research Studentships from the Science and Engineering Research Council (to N.K.D. and E.N.M.) and from the British Technology Group (to N. McK.).

REFERENCES

Abeles, R. H. (1979) in *Vitamin B₁₂* (Zagalak, B. & Friedrich, W., eds.), pp. 373–388. W. de Gruyter, Berlin and New York.
 Babior, B. M. & Krouwer, J. S. (1979) *CRC Crit. Rev. Biochem.* **6**, 35–97.
 Biggen, M. D., Gibson, T. J. & Hong, G. F. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 3963–3965.
 Brown, C. E., Shemin, D. & Katz, J. J. (1973) *J. Biol. Chem.* **248**, 8015–8023.
 Buck, D., Spencer, M. E. & Guest, J. R. (1985) *Biochemistry* **24**, 6245–6252.
 Chang, J.-Y. (1983) *Methods Enzymol.* **91**, 455–466.
 Chou, P. Y. & Fasman, G. D. (1978) *Annu. Rev. Biochem.* **47**, 251–276.

Dieckgraefe, B. K., Seetharam, B., Banaszak, L., Leykam, J. F. & Alpers, D. H. (1988) *Proc. Natl. Acad. Sci. U.S.A.* **85**, 46–50.
 Dolphin, D. (1982) in *B₁₂* (Dolphin, D., ed.), vol. 2, pp. 502–546. John Wiley and Sons, New York.
 Dretzen, G., Bellard, M., Sassone-Corsi, P. & Chambon, P. (1981) *Anal. Biochem.* **112**, 295–298.
 Finke, R. G. & Hay, B. P. (1984) *Inorg. Chem.* **23**, 3041–3046.
 Flinta, C., Persson, B., Jörnvall, H. & von Heijne, G. (1986) *Eur. J. Biochem.* **154**, 193–196.
 Francalanci, F., Davis, N. K., Fuller, J. Q., Murfitt, D. M. & Leadlay, P. F. (1986) *Biochem. J.* **236**, 489–494.
 Frank, R., Heikens, W., Heisterberg-Moustis, G. & Blocker, H. (1983) *Nucleic Acids Res.* **11**, 4365–4377.
 Grosjean, H. & Fiers, W. (1982) *Gene* **18**, 199–209.
 Hanahan, D. (1983) *J. Biol. Chem.* **166**, 557–580.
 Kellermeyer, R. W., Allen, S. H. G., Stjernholm, R. & Wood, H. G. (1964) *J. Biol. Chem.* **239**, 2562–2569.
 Kyte, J. & Doolittle, R. F. (1982) *J. Mol. Biol.* **157**, 105–132.
 Ledley, F. D., Lumetta, M., Nguyen, P. N., Kolhouse, F. J. & Allen, R. H. (1988) *Proc. Natl. Acad. Sci. U.S.A.* **85**, 3518–3526.
 Maas, R. (1983) *Plasmid* **10**, 296–298.
 Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor.
 Marsh, E. N., Leadlay, P. F. & Evans, P. R. (1988) *J. Mol. Biol.* **200**, 421–422.
 Matthes, H. W. D., Zenke, W. M., Grundstrom, T., Staub, A., Wintzerith, M. & Chambon, P. (1984) *EMBO J.* **3**, 801–805.
 Messing, J. (1983) *Methods Enzymol.* **101**, 20–78.
 Murtif, V. L., Bahler, C. R. & Samols, D. (1985) *Proc. Natl. Acad. Sci. U.S.A.* **82**, 5617–5621.
 Needleman, S. J. & Wunsch, C. D. (1970) *J. Mol. Biol.* **48**, 443–453.
 Overath, P., Stadtman, E. R., Kellerman, G. M. & Lynen, F. (1962) *Biochem. Z.* **336**, 77–98.
 Remington, S., Wiegand, G. & Huber, R. (1982) *J. Mol. Biol.* **158**, 111–152.
 Rétey, J. (1982) in *B₁₂* (Dolphin, D., ed.), vol. 2, pp. 357–379. John Wiley and Sons, New York.
 Samols, D., Thornton, C. G., Murtif, V. L., Kumar, G. K., Haase, F. C. & Wood, H. G. (1988) *J. Biol. Chem.* **263**, 6461–6464.
 Sanger, F., Nicklin, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463–5467.
 Schwartz, R. M. & Dayhoff, M. O. (1979) in *Atlas of Protein Sequencing and Structure* (Dayhoff, M. O., ed.), vol. 5, suppl. 3, pp. 345–362. National Biomedical Research Foundation, Washington.
 Staden, R. (1982) *Nucleic Acids Res.* **10**, 4731–4751.
 Staden, R. (1984) *Nucleic Acids Res.* **12**, 551–567.
 Suggs, S. V., Wallace, R. B., Hirose, T., Kawashima, E. & Itakura, K. (1982) *Proc. Natl. Acad. Sci. U.S.A.* **78**, 6613–6617.
 Vieira, J. & Messing, J. (1982) *Gene* **19**, 259–268.
 Wiegand, G., Remington, S., Deisenhofer, J. & Huber, R. (1984) *J. Mol. Biol.* **158**, 205–219.
 Wölfe, K., Michenfelder, M., König, A., Hull, W. E. & Rétey, J. (1986) *Eur. J. Biochem.* **156**, 545–554.
 Wollowitz, S. & Halpern, J. (1984) *J. Am. Chem. Soc.* **106**, 8319–8321.
 Wood, H. G. & Zwolinski, G. K. (1976) *CRC Crit. Rev. Biochem.* **4**, 47–122.
 Zagalak, B., Rétey, J. & Sund, H. (1974) *Eur. J. Biochem.* **44**, 529–534.
 Zhang, H., Scholl, R., Browse, J. & Somerville, C. (1988) *Nucleic Acids Res.* **16**, 1220–1228.