Nucleotide Sequence of the *uhp* Region of *Escherichia coli*

MARY JANE FRIEDRICH AND ROBERT J. KADNER*

Department of Microbiology and the Molecular Biology Institute, School of Medicine, The University of Virginia, Charlottesville, Virginia 22908

Received 16 March 1987/Accepted 20 May 1987

The Escherichia coli uhp region encodes the transport system that mediates the uptake of a number of sugar phosphates as well as the regulatory components that are responsible for induction of this transport system by external glucose 6-phosphate. Four uhp genes have been identified by analysis of the complementation behavior and polypeptide coding capacity of plasmids carrying subcloned regions or transposon insertions. The nucleotide sequence of a 6.5-kilobase segment that contains the 3' end of the *ilvBN* operon and the entire *uhp* region was determined. Four open reading frames were identified in the locations expected for the various uhp genes; all were oriented in the same direction, counterclockwise relative to the genetic map. The properties of the polypeptides predicted from the nucleotide sequence were consistent with their observed features. The 196-amino-acid UhpA polypeptide has the composition characteristic of a soluble protein and bears homology to the DNA-binding regions of many regulatory activators and repressors. The 518-amino-acid UhpB and the 199-amino-acid UhpC regulatory proteins contain substantial segments of hydrophobic character. Similarly, the 463-amino-acid UhpT transporter is a hydrophobic protein with numerous potential transmembrane segments. The UhpC regulatory protein has substantial sequence homology to part of UhpT, suggesting that this regulatory protein might have evolved by duplication of the gene for the transporter and that its role in transmembrane signaling may involve sugar-phosphate-binding sites and transmembrane orientations similar to those of the transport protein.

Expression of the *Escherichia coli* sugar phosphate transport system is induced by external glucose 6-phosphate, but is unaffected by glucose 6-phosphate generated internally from phosphorylation of exogenous glucose (33). Previous studies using *uhpT-lac* operon fusions showed that induction of the transporter gene *uhpT* occurred normally even in strains lacking detectable glucose 6-phosphate uptake activity (25). The characteristics of Uhp expression suggested that regulation involves transmembrane signaling, in which a membrane-bound regulatory protein serves as the receptor for external glucose 6-phosphate and triggers events that result in elevated transcription of the *uhpT* gene (32).

All genes specifically involved in the production of the Uhp transport system are linked in the *uhp* region at min 82.1 of the *E. coli* genetic map (12, 13). The accompanying paper describes the location of the four *uhp* genes and their requirement for Uhp expression (32). Of the three regulatory genes, *uhpA* appears to encode a positive activator of *uhpT* transcription, as suggested by its absolute requirement and the constitutive expression of the UhpT transporter when *uhpA* is present at elevated gene dosage. Two other genes, *uhpB* and *uhpC*, encode proteins that are required for *uhpT* expression except when *uhpA* is present in multicopy plasmids. Complementation results also suggested that the UhpT transporter is comprised of only a single polypeptide.

This paper communicates the nucleotide sequence of the *uhp* region carried on a 6.5-kilobase (kb) *Hind*III-*Eco*RV fragment from the Clarke-Carbon plasmid pLC17-47 (26). The location of the open reading frames correlated well with the location of the genes defined by subclones and transposon insertions (32). Homologies of portions of the polypeptides deduced from the nucleotide sequence support a model for regulation through transmembrane activation of a transcriptional activator.

MATERIALS AND METHODS

Determination of nucleotide sequence. DNA sequencing was performed by the dideoxy chain termination method of Sanger et al. (23), using $[\alpha^{-35}S]$ thio-dCTP (5). Singlestranded DNA templates were derivatives of the phage vectors M13 mp18 and M13 mp19 (18) carrying restriction fragments from the *uhp* region of plasmid pRJK10 (32). These included small restriction fragments generated by cleavage with Sau3A, TaqI, HpaII, or PvuII. Larger fragments were generated by isolation and ligation of the 1.58-kb HpaI, the 1.96-kb ClaI, the 4.2-kb ClaI-2-BamHI, the 2.4-kb SphI-BamHI, the 790-base-pair SphI-Bg/II, and the four EcoRV fragments in both orientations in the vectors cut with appropriate restriction enzymes.

Nested deletions extending into the larger DNA fragments from the end proximal to the universal primer-binding site on the vector were generated as described by Dale et al. (7). In this method, single-stranded viral DNA is converted to linear form by cleavage with an appropriate restriction enzyme (*HindIII* or EcoRI) after hybridization to the template of an oligonucleotide complementary to the sequences around the restriction site. Treatment of the linear DNA with T4 DNA polymerase results in generation of deletions into the insert by means of the enzyme's 3' to 5' exonuclease activity. Recircularization is then accomplished by addition of a short 3' homopolymer tail, followed by annealing to an oligonucleotide complementary to both ends and subsequent ligation of the juxtaposed ends.

Gel readings of areas of high G+C content often resulted in sequence anomalies (compression). More reliable readings through these areas were obtained by substitution of 7-deazadeoxyguanosine 5'-triphosphate in place of dGTP (3).

The management of sequence information and subsequent analysis of compiled data was effected by the DBUTIL programs of Staden (29). Hydropathy profiles were deter-

^{*} Corresponding author.



FIG. 1. Restriction map of the *uhp* region and sequencing strategy. The top line presents the location of restriction enzyme cleavage sites that were used for determination of the nucleotide sequence or for construction of subcloned plasmids. Multiple sites for a single enzyme are numbered from left to right. Numbering begins at the *ClaI*-1 site. The sequencing strategy diagrammed in the central section gives the length and direction of each fragment whose sequence was determined. Lines with filled arrowheads and beginning with a vertical line represent restriction fragments, whereas lines with open arrowheads with a circle represent deletions generated by the action of T4 DNA polymerase. At the bottom is presented schematically the location of the open reading frames and their gene assignments, as described in the text.

mined by using the parameters of Kyte and Doolittle (15) or Engelman et al. (8). Comparisons of amino acid sequences used the FASTP program of Lipman and Pearson (16).

Reagents. Restriction endonucleases and T4 DNA ligase were purchased from New England Biolabs, Inc., or Boehringer-Mannheim Biochemicals and were employed under the conditions recommended by their manufacturers. The DNA sequencing kits and $[\alpha^{-35}S]$ thio-dCTP were obtained from Amersham Corp. The T4 DNA polymerase and oligonucleotides used for generation of deletions were obtained from International Biotechnologies, Inc.

RESULTS AND DISCUSSION

Nucleotide sequence of the *uhp* region. The nucleotide sequence of the 6.5-kb *HindIII-Eco*RV-6 DNA fragment carried in plasmid pRJK10 was determined by means of the sequencing strategy shown in Fig. 1. Both defined restriction fragments and random deletions generated by the action of T4 DNA polymerase on single-stranded phage M13 templates were analyzed. Numbering of the nucleotides begins at the *ClaI-1* site. Except for a 124-base-pair stretch (nucleotides 1956 to 2080), sequence data were obtained for both strands, with overlaps covering all restriction sites. The assignment of each nucleotide was based on an average of 6.7 gel readings.

Identification of open reading frames in the *uhp* region. The genetic studies reported in the accompanying paper indicated the existence of four *uhp* genes (32). Four translational reading frames of appropriate sizes were deduced from the nucleotide sequence (Fig. 2), in locations consistent with the *uhp* gene boundaries defined by transposon insertions and subclones. The coding region of the *ilvBN* operon ends at nucleotide 84 and is followed by a typical Rho-independent transcription termination sequence, as described by Wek et al. (31) and Friden et al. (9).

Just past this termination sequence, the putative uhpA reading frame extends from nucleotide 163 to nucleotide 750, encoding a 196-amino-acid polypeptide of M_r 20,893. A potential Shine-Dalgarno ribosome-binding sequence (AGGA) (28) is located 7 nucleotides upstream from the start of this reading frame. No other potential initiation site in this or another reading frame (at nucleotides 302, 304, 310, 312, 331, 379, 391, or 479) was preceded by this collection of characteristic expression signals, although amino-terminal sequence information is necessary to define the actual start site.

The uhpA gene ends with the sequence TGATGA at nucleotide 751; the ATG in this sequence is at the beginning of the long uhpB reading frame between nucleotides 753 and 2306. This reading frame is preceded by a possible Shine-Dalgarno sequence (GATGG) and encodes a 518-amino-acid polypeptide of M_r 58,642. Other possible initiation sites for this gene (nucleotides 864, 870, or 906) are not preceded by likely Shine-Dalgarno sites. It is possible that the translation of uhpA and uhpB is coupled, although there is no direct evidence for this point. The location of the 3' end of the *uhpB* reading frame is in agreement with the end of the gene defined by transposon insertions. Insertion 4 in uhpB is located between nucleotides 1956 (the ClaI-2 site) and 2341 (an AhaII site), whereas insertion 28, which lies outside uhpB, is located between nucleotides 2341 and 2459 (the EcoRV-4) site (32).

A third open reading frame is found between nucleotides 2457 and 3272. However, the first initiation codon is the ATG at nucleotide 2616, which would predict a 219-aminoacid polypeptide product with M_r 23,790. This initiation codon is preceded by a very poor ribosome-binding sequence (GA), and the predicted polypeptide is substantially larger than the observed M_r 20,000 UhpC product. It is more likely that the *uhpC* coding sequence is initiated at the GTG codon at nucleotide 2676. Translation from this site would

IDKLEDVVKVQRNQSDPTMFNKIAVFFQ ATCGATAAGCTGGAAGATGTGCGTGGAAGTGCGAGGCGTAATCGGCGCGTGTTTAAGAAGATCGCGGGTGTTTTTCAGTAACCACGCGCTTGAACAACATCGCGCTTAATCG 50 100
MITVALIDDHLIVRSGFAQLLGCLGCC MITTVALIDDHLIVRSGFAQLLGCCGCCAGGACAAGACCATGATCACCGTTGCCCCTCATGGCCCCCCCGGCCTTGCCGCAGCGCGGGGCTGGGAACCTGAT 150 200 200 200 200 200 200 200 200 200 2
L Q V V A E F G S G R E A L A G L P G R G V Q V C I C D I S M P D I S G L E L L TIGCAGGTAGTTGCCGAGTTGGGTCGGGGGCGCGGGGGGGGG
S Q L P K G M À T I M L S V H D S P A L V E Q A L N A G A R G F L S K R C S P D AGCCAGCTGCCGANAGGTATGGCGACGATTATGCTCTCCGGTGAGCAGGCGGCGTTAALGGCGGGGCGACGCGGCGTTTCTTTCCAAACGCTGTAGCCCGGAT * 400 * 450 *
ELIAAVHT VAT GGCYLT PDIAIKLASGR QDPLT KRER QVA GAACTCATTGCTGCGGTGCATACGGTTGCCACGGGCGGCTGTTATCTGACGCCGGGATATTGCCATTGCCATCGGCTCGGCGCGCGC
EKLAQGMAVKEIAAELGLSPKTVHVHRANLMEKLGVSNDV GAAAAACTGGCGCAAGGAATGGCGGTGGAAGGAGTGCCGCGACGTGCCACGGCCAATCGGCGCGCCAATCGGCGCGCAATCGGCGGCGCCAATCGGCGCGCGC
UhpB MKTLFSRLITVIACFFIFSAAWFCLWSISL ELARRMFDGW* GAGCTGGCGCCGCCATGTTTGATGGCTGGTGATGAAGACGTTGTTCCCCGCCTAATTACCGTCTTGCCTGCTTTTTATCTTCTCTGCCGCGATGTTTTGCCTGGGGGGATATCAGCC
* 750 * * 800 * * H L V E R P D M A V L L F P F G L R L G L M L Q C P R G Y W P V L L G A E W L L TGCATCTGGTTGAGCGCCCCGATATGGCGGGGGGGGGGG
I Y W L T Q A V G L T H F P L L M I G S L L T L L P V A L I S R Y R H Q R D W R TGATTACTGGCTAACGCAGGCGGTCGGTTTAACCAATTTCCGTAATGACGGTACTTACT
T L L L Q G A A L T A A A L L Q S L P W L W H G K E S W N A L L L T L T G G L T GCACCTTGCTGTTAACGGGGGGGGGTTAACGGGGGGGGGG
L A P I C L V F W H Y L A N N T W L P L G P S L V S Q P I N W R G R H L V W Y L CGCTGGGCCCCGATATGTCTGGGTTCTGGCACTATCTCCGGCCATCGGTCGG
L L F V I S L W L Q L G L P D E L S R F T P F C L A L P I I A L A W H Y G W Q G TGCTGCTGCTGTTGTTATCAGTCTCGGCCTCGGCGCTGGGAGTGCCGGAGGGAG
A L I A T L M N A I A L I A S Q T W R D H P V D L L L S L L V Q S L T G L L L G GGGGGGTGATGGGGGGGGGGGGGGGGGGGGGGGGGGG
A G I Q R L R E L N Q S L Q K E L A R N Q H L A E R L L E T E E S V R R D V A R GCGCTGGCATCCAGCGGTTGCGTGAACTTAACCAGTGGCGCGCAAAAGGAACTGGCGGCGCAATCAGCATCTGGCTGAAACGGAAACGGAAGAGAGGGGGGCGCGGGGGGG 1600 1650
ELHDDIGQTITAIRTQAGIVQRLAADRRQREAERAAHRTT GTGAGCTGCATGATGATGATGATGAGCGAGCCATCATCGCTACTCGAGCGGGGGGGG
I A G R L R R V R R L L G R L R P R Q L D D L T L E Q A I R S L M R E M E L E G CTATCGCTGGGCGTTTACGACGGGGGGGGGGGGGGGGG
R G I V S H L E W R I D E S A L S E N Q R V T L F R V C Q E G L N N I V K H A D GGCGCGGGTATTGTCAGCCATCTCGAATGGCGAATCGATGATGAGCGTTAAGCGAAAACATGCGG 1950 2000
A S A V T L Q G H Q Q D E R L M L V I E D D G S G L P A G S G N K V L A S P E C ATGCCAGCGCGGCGCCCCCGCAAGCAGGATGAACGGTTGATGCTGGTTATGCAAGGCAGGC
A T R N G A G W H I T H F L S A R H A C Q R F S T S T L C L R F D D V A V S E S GCGCGACGCGTAACGGCGCTGGGTGGCACATTACACATTACCTGTCTGCGCGGCGCGCGC
A C R C A I N D * GCGCCTGCCGATGCCCATTAATGACATATGAAATTGAAATTGATGCCCGCTATCGCCTACTGGCGCGGCATATTCTGCTGACCATCTGGCTGG
ANANGTITTAACGCCGCCGTACCAGANATCCTTGCTAACGCCGTGCTCACCGATATCGGCCTGTTAGCGACCCTGTTTACATTACCTATGGCTGTTGAAGATTGCCCGCG 2450 UDpC W G V C R A L G A
ATTGTCAGCGATCGCTCAAAAGCCCGTTATTTTATGGGGATAGGGCTTATCGCCACGGGCATTATCAACATTCTGTTTGGCTTCTCGACGTCGCCATGGGCGTTTGCCCGTGCTCTGGGGG 2550 2600
E R L F P G L G F T G V C A S V N G L V F T Y R A R R W W A L W N T A H N V G G CTGAACGCCTTTTTCCAGGGGTTCACCGGTGTGTGGGGGCGTTGTGGGGGGTTGGGGGGGG
A L I F I V M A A A A L H Y G W R A G M M I A G C M A I V V G I F L C W R L K D GCGCACTCATTGCCATGGCAGCGGCGGCGGCGGCGGCGGGCG
R P Q A L G L P A V G E W R H D A L E I A Q Q Q E G A G L T R K E I L T K Y V L ATCGCCCGCAGGCGTTAGGTTTACCGGCGGTGAATGGCGACACGACGCGCGGGAATGGCGCGGGAATGGCGGGGGGGG
L N P Y I W L L S F C Y V L V Y V V R A A I N D W G N L Y M S E T L G V D L V T TGCTGAATCCGGTATATCTGGCTGCTTTGGTTTTGGTATGTGGTGCTGGGGGCGATCAACGACTGGGGGCAATTGGTATATGTCCGAGAACTGGGGGCGATCTAGGCGA 3050
ANTAVTMFELGGFIGALVAGWGSDKLFNGNRGPMNLIFAA CGGCGAATACGGCAGTGACGATGTTGAACTGGGCGGATTATCGGTGGGCGGGTGGGGCTCGGACAATTGTTTACGGCAACCGAGGGCCGATGAATTGATTTCGCCG 3200
G I L L S V G S L C CCGGAATTITGCTTTCACTGGGCTCCCTGTGGTGGTGGTGCAGCCAGCGGGCAACCTGCTTCTTCACCATTGGTTTTTTGTCTTTGGCCCACAGATGTTAATCGGTAT 3250 3250

FIG. 2. Nucleotide sequence of the uhp region. The nucleotide sequence extending for 5,400 base pairs from the Clal-1 site is presented. The predicted amino acid sequences for the uhp open reading frames are indicated in one-letter code.

GCCGCCGCCAGAGTGTTCCCAC	AAAGAGGCGGCAGGGGGGGGGGGGGGGGGGGGGGGGGGG	GGGTTTGTCGGCTTGTTTGCTT	TCTGGGGGGCGTCGCTTGCTGG	TTGGCCGCTGGCGAAAGTACTCGATAC	CTG
*	3400	*	* 3450	•	*
GCACTGGAGCGGATTTTTTGTG	GTTATCTCTATCGCCGCCGGGAT:	TCCGCACTGCTGTTACTGCCCT	TTTTGAACGCCCAGACACCGCG	CGAAGCGTGATGCATCTCACCTTTTCA	CTT
3500	*	* 35	50 *	* 3	600
					Uhp
CATATCCGGCAAAACTAAGAAA	TTTTCCAGGTTTTGCCTGGACGC	2650	CATTITIACAATGCATGCCTCA	3700	AIG *
	-	3630 -	-	3700	
LAFLHOV	REPTLDLP		FEPEMOS	Y L V V P I G Y L	T
CTGGCTTTCTTAAACCAGGTTC	CANCEGACCTGGACCTTCCG	TCGAAGTGCGCCCAAAATGTG	TTCANACCGTTCATGCAATCO	TACCTGGTGGTCTTTATCGGCTACCTG	ACG
*	3750 *	*	3800	*	*
MYLIRKNI	F N I A Q N D M	ISTYGLSM	TQLGMIG	LGFSITYGV	G
ATGTACCTGATTOGCAAGAACT	TTANCATCGCGCAGAACGATATG	ATTTCGACCTACGGGTTGAGCAT	GACGCAGCTGGGGATGATCGGC	CTGGGTTTCTCCATCACTTATGGCGTG	GGT
3850 *	*	3900	•	* 3950	*
KTLVSYY.	A D G K N T K Q	FLPFMLIL	SAICMLG	FSASMGSGS	V
ANACOUTOGTITCCTACTACG	CCGACGGCAAAAACACCAAACAA	TTCCTGCCGTTCATGCTGATCCT	CTCTGCTATTTGTATGCTGGGC	TTCAGTGCCAGTATGGGCAGCGGCTCG	GTT
-	4000	•	* 4050	•	•
			 .		τ.
ACCTETTOTEATEATECCT	TTACCOTTANCCCCTTTTTC	A S I G G S C S	TACTCCACCATCACCAAATGG	ACCCCCCCCCCTCAAACCCCGGGACATTC	cic
4100	*	* 41	50 *	* 4	200
GFWWISE	N L G G A G A A	GVALFGAN	YLFDGHV	IGMFIFPSI	I
GETTTCTGGAATATTTCTCACA	ACCTTOGCGGTGCAGGCGCAGCA	GTGTGGCGCTGTTCGGGGCAAN	TTACCTGTTCGATGGCCATGTC	ATCOGCATOTTATCTTCCCGTCGATT	ATC
+	*	4250 *	+	4300	*
ALIVGFI	GLRYGSDS	PESYGLGK	A E E L F G E	EISEEDKET	E
COLICENTERIO	OCCTOCOTTACGCAGCGACTCC	CGGAATCTTATGGCCTCGGCAA	NGCTGAAGAACTGTTCGGCGAG	GAGATCAGOGAAGAGGACAAAGAGACA	GAA
*	4350 *	CGGAATCTTATGGCCTCGGCAA *	AGCTGAAGAACTGTTCGGCGAG 4400	GAGATCAGCGAAGAGGACAAAGAGACA *	GAA *
	4350 *	CGGAATCTTATGGCCTCGGCAA	AGCTGAAGAACTGTTCGGCGAG 4400	CAGATCACCGAAGAGGACAAAGAGACA	GAA *
S T D N T K W	Q I F V E Y V L	K N K V I W L L	ACCTGAAGAACTGTTCGGCGAG 4400 C F A N I F L	GAGATCHCCGAAGAGGACAAAGAGACA * Y V V R I G I D Q	GAA * W
S T D M T K W TCTACCEATATCACCAAGTOOC	Q I F V E Y V L BEATCTTTGTTGAGTAGTGCGG	K N K V I W L L	AGCTGANGANCTGTTCGGCGAG 4400 C F A N I F L GTGCTTCGCCANCATTTCCTC	CHARTCAGCEANGAGENCAANGAGACA * Y V V R I G I D Q TATGTGGTAGGTATGGTATGGTACGA	GAA W TGG
S T D M T K W TCTACCGATATGACCAMSTGGC 4450 *	Q I F V E Y V L MGATCTTTCTTGAGTATGTGCTG.	K N K V I W L L AMAAACAAAGTGATCTGGCTGCT 4500	AGCTGAAGAACTGTTCGGCGAG 4400 C F A N I F L GTGCTTCGCCAACATTTTCCTC #	CACATCROCCAAGAGACAAAGAGACA * Y V V R I G I D Q TATGTGGTACGTATTGGTATGGACCAG * 4550	GAA ¥ TGG
S T D N T K W TCTACCCATATCACCAAGTCCC 4450 * S T V Y A F O	CONCONTINUES ACCOUNT OF A	K N K V I W L L ANAACAAAGGATCGCGCGCT 4500	AGCTGAAGAACTGTTCGGCGAG 4400 C F A N I F L GTGCTTCGCCAACAATTTTCCTC T	TATGTGGTACGACHAGAGACAAAGAGACAA Y V V R I G I D Q TATGTGGTACGTATGGTATGGACCAG * 4550 T L L W G W L S D	GAA W TCG
* S T D M T K W TCTACCGATATGACCAAGTGGG 4450 * S T V T A F Q TCAACCGTATACGGGTTCCAGG	CCCCCCCTACCCCACCCACCCACCCA 4350 * Q I F V E Y V L MCATCATACTORATATGACCAC * E L K L S K A V	K N K V I W L L ANAAACAANGTGATCTGGCTGCT 4500 A I Q G F T L F 20GATCAAGGCTTTACGCTGTT	AGCTGANGANCTGTTCGGCGAG 4400 C F A N I F L GTGCTTCGCCANCATTTCCTC * E A G A L V G TGANGCTGTGCGCTGGTGGGT	CHARTCHOCCHARGHORACHANGHACHA Y V R I G I D Q TATGTGGTACGTATTGGTATCGACCAG * 4550 T L L W G W L S D ACCTOCTGGGCGCGCGCTGCTCTCAC	GAA W TGG L CTG
S T D M T K W TCTACCCANATCACCAMOTOCC 4450 * S T V T A F Q TCAACCGTATACCCCTTCCACG	CONCENTRATION ACCOUNT	K N K V I W L L ANALACANGTOATCOGCOCC 4500 A I Q G F T L F SCGATTCAGGOCTTIAGGCOTT	AGCTGANGANCTGTTCGGCGAG 4400 C F A N I F L GTGCTTCGCCANCATTTTCCC * E A G A L V G TGANGCTGGTCGCCTCGT * 4650	CACATCHOCCAMERICANGACAM Y V R I G I D Q TATGTGCTACTATTGCTATCGACCAG * 4550 T L L W G W L S D ACGCTGCTGCGCCTGCTCTCTGAC	GAA W TGG ± L CTG
S T D N T K W TCTACCEATATGACCAMORGEC 4450 * S T V T A F Q TCANCOURTAGECCATTOCAGE	GCTGOCTACGCACCACTCC 4350 Q I F V E Y V L MARTCHTRETGAGTARGEGEGE * E L K L S K A V **	K N K V I W L L AMANACANATGATCTGGCTGGT 4500 A I Q G F T L F SCGATTCAGGCGTTT	AGCTGANGANCTGTTCGGCGGAG 4400 C F A N I F L GTGCTTCGCCANCATTTTCCTC E A G A L V G TGANGCTGGTGCCCTGGTGGGT * 4650	CACATCROCALGAGEAGEAGEAGEAGEAGEAGEAGEAGEAGEAGEAGEAGE	GAA ¥ TGG * L CTG
S T D M T K W TCTACCEATATEACCAAGTEGC 4450 S T V T A F Q TCAACCETATACCEGTOCCAGE	ACCTOCOTTACOCASCASCACCA 4350 0 I F V E Y V L MARCTITOTTCASTATCTCCTC * E L K L S K A V MACTCAAACCTCTAAACCGGTC 4600 V A C I A L A L	K N K V I W L L NAMACANASTGATCTGGCTGGT 4500 A I Q G F T L F SCGATTCHOGGCTTTMCGCTGTT I I A T L G V Y	AGCTGAAGAACTGTTCGGCGAG 4400 C F A N I F L GTGCTTCGCCAACATTTTCCTC * E A G A L V G TGAAGCTGGTGCGCTGGTGGGT 4650 Q H A S N E Y	CACATCHOCCAMERICANGACAM Y V R I G I D Q TATGTOGTACCTATTCGTATCGACCAG * 4550 T L L W G W L S D ACCOCTOCTOGCCGCTGCTCTCAC I Y L A S L F A L	GAA * TGG * L CTG *
S T D N T K W TCTACCEATATGACTAMOSTOC 4450 S T V Y A F Q TCANCECTATACCETTCCACE A H G R R G L CCANCECTCCCCTCCCCTCC	Q I F V E Y V L AGAICTITIGTIGAGIANCIGUE E L K L S K A V INCEGNACICICIANCOCIGUE CALL S K A V INCEGNACICICICANCOCIGUE V A C I A L A L TODOCTOCATCOCCTOC	K N K V I W L L ANALACANGTCATCGCGCGC 4500 A I Q G F T L F SCGATTCAGGCCTTAGGCCGTT T I A T L G V Y ATTATCGCCACGCCGCGGTGTA	AGCTGANGANCTGTTCCGCGAG 4400 C F A N I F L GTGCTTCGCCANCATTTTCCCC * E A G A L V G TGANGCTGGTGCCCTGGTGGGG * 4650 Q H A S N E Y TCANCATGCCGTANGGANTAS	CACATCACCTACEACEACEACEACEACEACEACEACEACEACEACEACEA	GAA * TGG * L CTG G GGT
S T D M T K W TCTACCEATATGACCAMETGOC 4450 * S T V T A F Q TCAACCETATACCEATTCAAC * A W G R R G L CCEAACCETCECCTGECCTGEC 4700	CONCENTRATION CONCENTRATION 4350 0 I F V E Y V L MARCHINGTANGTANGTACTAC E L K L S K A V MARCHANCTONTANGCOSTO 4600 V A C I A L A L TRECOTOCATOGCACTOGCOCTOS	ECCANTETTATGECETCOGEAN K N K V I W L L NANACANAGEGATETGECTECT 4500 A I Q G F T L F SCGATECMOSOCITIACGEGETT I I A T L G V Y MITATEGECEMECCTEGATETA 4	AGCTGAAGAACTGTTCGGCGAG 4400 C F A N I F L GTGCTTCGCCAACATTTTCCTC E A G A L V G TGAAGCTGGTGCGCTGGTGGGT 4650 Q H A S N E Y TCAACATGCCAGTAAGGAATAT 50	CACATCHOCCAMERICANGACAM Y V V R I G I D Q TATGTGGTACGTATTGGTATCGACCAG 4550 T L L W G W L S D ACGCTOCTGGGGCTGCTGCGCCTCTGAC I Y L A S L F A L I Y L A S L F A L	GAA TGG L CTG G GGT 800
S T D M T K W TCTACCEATATEACCAMETECC 4450 S T V T A F Q TCAACCETATEACCETTCCMEE A W G R R G L CCEAACEGTCCCCCTGCCTGC 4700	ACCTOCOTTACOCCACCACTCC 4350 O I F V E Y V L MARCTITOTTCAGTATCTCCTC E L K L S K A V MACTCAAACCGCTCAAACCGGTC 4600 V A C I A L A L TOGCCTOCATCOCCCTGGCCCTC.	K N K V I W L L ANAMCANAGTCATCTGGCTGCT 4500 A I Q G F 7 L F SCGATTCHOGGCTTTAGGCTGTT I I A T L G V Y ATTATCGCCACGCTGGGTGTGTA 47	AGCTGANGANCTGTTCGGCGAG 4400 C F A N I F L GTGCTTCGCCANCATTTTCCTC * E A G A L V G TGANGCTGGTGGCGCGCTGGTGG 4650 Q H A S N E Y TCANCATGCCAGTANCGANTAJ 50 *	CACATCHOCCAME AGACAMAGACAA Y V V R I G I D Q TATCTOGTACCTATTGCTATCGACCAG * 4550 T L L W G W L S D ACCCTOTTGCGCTGCGCCTGCTGCAC I Y L A S L F A L PATCTATCTGCCTTGCGTTGCGTTGCGTTGCGTTGCGTTG	GAA W TGG L CTG G GGT 800
S T D N T K W TCTACCENTATCACCAMOTOCC 4450 S T V Y A P Q TCANCCETATACCCETTCCMOC COMPCETTCCCCTCCCCC 4700 F L V F C P Q	CONCECTION CONCENTRATION 4350 Q I F V E Y V L MEATCHTHETTENETANGTANGTOCTE E L K L S K A V MACTENANCTECTANGEOGTE 4600 V A C I A L A L TODOCTOCATEOCOCTOCCTEC L L I G V A A V	K N K V I W L L AMAAAAAAGGAACGGGCGCC 4500 A I Q G F T L F SCGATTCAGGGCTTTACGCGCGT I I A T L G V Y ATTATCGCCACGCCGGGGGGGTA 477	AGCTGANGANCTGTTCGGCGAG 4400 C F A N I F L GTGCTTCGCCANCATTTTCCTC # A G A L V G TGANGCTGGTGCCCTGGTGGGT * 4650 Q H A S N E Y TCANCATGCCCMGTANCGATAN 50 G A A D G I K	CACATCHOCCAMERICANCANGACAA Y V V R I G I D Q TANGTGGTANGGTATTGGTATCGACCAG 4550 T L L W G W L S D ACGCTOCTGGGGGCTGGCTCTCTGAC I Y L A S L F A L IATCTANCTGGCTTCTCTCTTGCGTTG 4 G T F A Y L I G D	GAA W TGG L CTG GGT 800 S
S T D M T K W TCTACCCATATCACCAMGTCCC 4450 * S T V T A P Q TCAACCGTATACGCGTTCCAGG A W G R R G L CCCAACCGTATACGCGTCCTGC 4700 P L V F G P Q TTCCTGCTCTTGCCCCCAM	CONCECTION CONCENTRATION 4350 Q I F V E Y V L MARCENTACTORENTATORECORE 4600 V A C I A L A L MARCECARCECTORECOREC 4600 V A C I A L A L MARCECARCECTORECOREC L L I C V A V MARTENTERTECTORECORECORECTOR	K N K V I W L L NANACANACTGATCTGGCTGCT 4500 A I Q G F 7 L F SCGATTCAGGCCTTTAGGCGTGTT 1 I A 7 L G V Y MTTATCGCCAGCCTGGGTGTGTA 47 G F V P K K A 7 SCCTTTGTACCTAAAAAACGAT	AGCTGANGANCTGTTCGGCGAG 4400 C F A N I F L GTGCTTCGCCANCATTTTCCTC * E A G A L V G TGANGCTGGTGCGCTGGTGCGT 4650 Q H A S N E Y TCCANCATGCCGTATGGATATAT 50 G A A D G I K TGGCGCTGCCGATGGTATTATAT	CACATCHOCCAMERICANGACAM Y V V R I G I D Q TATGTGGTACGTATTGGTATCGACCAG * 4550 T L L W G W L S D ACCCTOCTGGGGCCTGGCTCTCTAC I Y L A S L F A L ANCTATCTGGCTTCTCTCTTTGCGTTG 4 G T F A Y L I G D MCGCCCTTTCCTTACCTGATTGGTGAC	GAA W TGG L CTG GGT 800 S SAGC
S T D N T K W TCTACCENTATGACTAMETACA 4450 S T V Y A P Q TCANCESTATACCESTTOCACE A N G R R C L CCANCESTCCCOTESCTOC 4700 P L V F C P Q TTCTGETCTTFGCCCCCCA	CONCECTION CONCENTION CONCENTI	CCGCAATCTTATGGCCTCGGCAA K M K V I W L L ANAAACAAAGTGATCTGGCTGCT 4500 A I Q G F T L F CCGATTCAGGCCTTTAGGCTGTT 1 I A T L G V Y ATTATCGCCACGCTCGGTGTGTA 47 G F V P K K A I 4550 *	AGCTGANGANCTGTTCGGCGAG 4400 C F A N I F L GTGCTTCGCCANCATTTTCCTC * E A G A L V G TGANGCTGGTGGCCGCTGTCGCCGCTGTGGCGCCGCCGATAGGAT 50 C H A S N E Y TCANCATGCCAGTANCGATATAJ 50 C A A D G I K TGGCGCTGCCGATGGTATTAJ	CACATCHOCCAMEROCACAMEROCAC Y V V R I G I D Q TATCTOGTACTATTOGTATCGACCAG * 4550 T L L W G W L S D ACOCTOCTGTOGCTGCCCCCTCTCAC I Y L A S L F A L CATCTATCTGCCTTCCCCTTGCCTTG * 4 G T F A Y L I G D ACOCCCTTCCTTACCTGATTGGTGAC 4900	GAA W TGG L CTG G GGGT 800 S SAGC *
S T D M T K W TCTACCEATATEACCAMATECC 4450 S T V T A P Q TCAACCETATACCEATTCAAC A W G R R G L CCEAACECTCECCETTCACCE 4700 P L V F G P Q TTCTGETCTTTGECCCCEAAT	CONCECTION CONCENTRATION CONCENTRATICO CONCE	K N K V I W L L NANACANAGTGATCTGGCTGCT 4500 A I Q G F T L F SCGATTCMGGCTTTACGCGTGTT I I A T L G V Y MTTATCGCCMGCGCGGGGGTGTTA 4 G F V P K K A I SCGTTTGTACCTANANAGCGAT 4850	AGCTGANGANCTGTTCGGCGAG 4400 C F A N I F L GTGCTTCGCCANCATTTTCCTC E A G A L V G TGANGCTGGTGCGTGGTGCGT 4650 Q H A S N E Y TCANCATGCCAGTANGGATTAN G A A D G I K TGGCGCTGCCGATGGTATTAN	CACATCHOCCAMERICANGACAAAGACAA Y V V R I G I D Q TATGTGGTACGTATTGGTATCGACCAC 4550 T L L W G N L S D ACGCTOCTGGGGCGCGCCTCTGCC I Y L A S L F A L I Y L A S L F A L G T F A Y L I G D GGCCCCTTTGCTTACCTGATTGGTGCC	GAA W TGG L CTG G GGGT 800 S AGC *
S T D M T K W TCTACCCATATCACCAMGTCCC 4450 * S T V T A F Q TCAACCGTATCACCGTTCCMCG A M G R R G L CCCAMCCGTCCCCCTCCCCCCC 7100 F L V F G F Q TTCCTCGTCTTFGCCCCCCCAM	ACCTOCOTTACOCCACCACTCC 4350 Q I F V E Y V L MARCTATACTACACTACTCCTC E L K L S K A V MARCTANACTCTTANCGGTG 4600 V A C I A L A L TROECTOCATCACCCTGCGCCGC L L I G V A A V TROTTANTOCTOTOCCCCTGTT M I A D G T P V	K M K V I W L L ANAMCANASTGATCTGGCTGGT 4500 A I Q G F T L F SCGATTCHOGGCTTTMCGCTGTT I I A T L G V Y ATTATCGCCACGCTGGGTGTGTA 47. G F V P K K A I SCTTTGTCCCAANAAAACGAT 4850 F G L T G W A G	AGCTGAAGAACTGTTCGGCGAG 4400 C F A N I F L GTGCTTCGCCAACATTTTCCTC * E A G A L V G TGAAGCTGGTGGCGCCGGCTGGGT GAAGCTGGCGCGCGGTGGGTGGGTGGG G A A D G I K GGGGGCCGCCGATGGTATTAAJ * T F A A L D I	CACATCHOCCAMERICANGACAMGAGACA Y V V R I G I D Q TATGTGGTACGTATTGGTATCGACCAG * 4550 T L L W G W L S D ACCGTOCTGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	GAA W TGG L CTG G GGT 800 S GGT 800 S CAGC
S T D N T K W TCTACCENTATEACEAMATGCC 4450 S T V Y A P Q TCANCCETATACGCCTTCCMGC CONCECTATACGCCTTCCMGC CONCECTATACGCCCTTCCMGC 4700 P L V F C P Q TTCCTGCTCTTMGCCCCCCCMA F A K L G L G TTTCCCAMCTTMGCTCTGCGA	CONCONTINUOUS ACCALCENT 4350 Q I F V E Y V L MEATCHTHETTENETENETENETE E L K L S K A V INCTENANCECTANCESTO 4600 V A C I A L A L TESECCTOCATESCECTOC L L I G V A A V TESTERATESCESTOSCECCEETE M I A D G 7 P V INCATESCESTOSCECCEETE	K W K V I W L L AMANACANAGGATCTGGCTGGT 4500 A I Q G F T L F SCGATTCAGGCCTTTACGCGTT I I A T L G V Y ATTATCGCCACGCTGGGGGGGTGTA 47 G F V P K K A I SCGTTTGTACCTANANAGGAT 4550 F G L T G W A G	AGCTGAAGAACTGTTCGGCGGAG 4400 C F A N I F L GTGCTTCGCCAACATTTTCCTC E A G A L V G TGAAGCTGGTCGCCCTGGTCGGT * 4650 Q H A S N E Y TCAACATGCCCGTAGTAATAGAATAT 50 G A A D G I K TGGCGCTGCCGATGGTATTAAJ * T F A A L D I CACCTTCGCCCGCGGGTATTAG	CACATCHOCCAMERICANGACAMGAGACA Y V V R I G I D Q TATGTGGTAGGTATGGTATGGACCAG 4550 T L L W G W L S D ACCOTOCTOGOGGCTGCCTCTCACC I Y L A S L F A L ACCOTOCTGGGGCTGCCTCTCTGCGTG 4 G T F A Y L I G D GCCCCCTTACCTTACCTGATGGTCCC 4900 A A I G C I C L M GCCCCCGATGGTTGTATCTGCCTGATG	GAA * W TGG * L CTG * GGT 800 S SAGC * A GGCG *
S T D M T K W TCTACCEATATGACCAMETGOC 4450 S T V T A P Q CCAMCCETATGACCEATECTOMO 4450 S T V T A P Q CCAMCCETATACCECTTOMO 4700 P L V F G P Q TTCCTGCTTTTCCCCCCAMAT P A K L G L G TTCCCAMETTAGCTCGCGAMAT	CONCONTRACTOR 4350 Q I F V E Y V L MARCENTERTREGENERE E L K L S K A V INACCONTRACTOR E L K L S K A V INACCONTRACTOR 4600 V A C I A L A L TRECCONTRACTOR CONCONTRACTOR 4000 V A C I A L A L TRECCONTRACTOR 1 L I G V A A V TRETENTION M I A D G T P V 1450 *	K W K V I W L L NANACANACTGATCTGGCTGGCT 4500 A I Q G F T L F SCGATCAGGGCTTTACGGCGTGT I I A T L G V Y MITATCGCCACGCGGGGGGGGGGGGGGGGGGGGGGGGGGGG	AGCTGANGANCTGTTCGGCGAG 4400 C F A N I F L GTGCTTCGCCANCATTTTCCTC * E A G A L V G C H A S N E Y TCANCATGCCGGTGGGTGGGT G A A D G I K TGGCGCTGCCCGATGGTATTANJ * T F A A L D I CACCTTCGCCGCGCGCGGTGATATC	CACATCHOCCAMERICANGACAAAGACAA Y V V R I G I D Q TATGTOGTACGTATTGGTATCGACCAG * 4550 T L L W G W L S D ACGCTOCTGGGGCTGGGCTGCTGTGAC I Y L A S L F A L AACCTATCTGGCTTGGCTGCCTGTGGC 4 G T F A Y L I G D MCGCACCTTGCTTACCTGATTGGTGAC 4900 A A I G C I C L M CCCOCCGATGGCCTGTGTACCTGCTGATG	GAA * W TGG CTG G GGT 800 S GGT 800 S S AGC * *
S T D N T K W TCTACCENTATCHCTANGTOC 4450 S T V Y A P Q TCANCCETATACCECTTOCACC COMMCCETATACCCCTTOCACC 4700 P L V F C P Q TTCCTCGTCTTTOCCCCCCACT P A K L C L C TTTCCCACTTACCTCCCCCACT	CONCECTINGUE ACCOLLAGE 4350 O I F V E Y V L MARCITATOTTOMOTATOTOCO E L K L S K A V MARCINANCTOTTOMOTATOTOCO V A C I A L A L TOGOCTOCATOCOCCOCTO L L I C V A A V MIA D C T P V V MIA	K W K V I W L L ANAMCANASTGATCTGGCTGCT 4500 A I Q G F 7 L F 5CGATTCHOGGCTTTMCGCTGTT 1 I A 7 L G V Y ATTATCGCCACGCTGGGTGTGTA 47. G F V P K K A I 5CTTTGTACCTANANAGCGAT 4850 F G L 7 G W A G FTCGGCCTTMCCGGCTGGGCMGG 1 Q Q L 7 V A *	AGCTGAAGAACTGTTCGGCGAG 4400 C F A N I F L GTGCTTCGCCAACATTTTCCTC * E A G A L V G TGAAGCTGGTGGCGCGGTGGGT 4650 Q H A S N E Y TCCAACATGCCAGTAACGAATAT 50 G A A D G I K TGGCGCTGCCGATGGTATTAAI * T F A A L D I CACCTTCGCCGCGCCTGGATATC 5000	CACATCHOCCAMERICANCANGACAA Y V V R I G I D Q TATGTOGTACCTATTGGTATCGACCAG * 4550 T L L W G W L S D ACCCTOCTOGGGGCTGGCTGCTGTGCGTG I Y L A S L F A L CATCTANCTGGCTGCTGCTGTTGGGTGG * 4 G T F A Y L I G D GGCACCTTTGCTTACTGATGGTGC 4900 A A I G C I C L M ************************************	GAA * W TGG CTGG GGT 8000 S GGT 8000 S AGCC * *
S T D M T K W TCTACCEATATEACCAMATGOC 4450 S T V T A P Q TCAMCCETATACCOCTTOCMOS A W G R R G L GCGAACCETCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CONCONTRACTOR ACCALCATION 4350 Q I F V E Y V L MARTCHTATTOR TANGTATOTOCTO E L K L S K A V MARTCHTANCTOR ALCOGTO 4600 V A C I A L A L TOUCTOCATOCCOCTOCOCTO M I A D G T V MONTRATTOCCOTOCOCTOCTOCTOT 8 K K I R R E K K COLANATOCCOCOCCOCALANAAAA	K W K V I W L L AMANACANAGGARCTGGCTGGT 4500 A I Q G F T L F SCGATTCHGGGCTTTAGGTGTT I I A T L G V Y MTTATCGCCAGGCCGGGTGTGTA 4 4 5 G F V P K K A I SCGTTTGTACCGAGGTGGGTAGG F F G L T G W A G F F C G L T G W A G F TCGGCCTTACCGGCTGGGCAGG	AGCTGANGANCTGTTCGGCGAG 4400 C F A N I F L GTGCTTCGCCANCATTTTCCTC E A G A L V G TGANGCTGGTGCGCTGGTCGGT 4650 Q H A S N E Y TCANCATGCCAGTANGGANTAT 50 G A A D G I K TGGCGCTGCCGGATGGATTATAN T F A A L D I CACCTTCGCCGGCGCGGATATC 5000	CACATCHOCCAMERICANCANGUARCA Y V V R I G I D Q TATGTGCTACGTATTGCTATCGACCAC 4550 T L L W G W L S D ACCTOCTGGCGCCTGCCTCTCTGC I Y L A S L F A L I Y L A S L F A L G T F A Y L I G D GCCACCTTRCCTTACCTGATTGCTGC A A I G C I C L M COCCCCATTGCTGTATCTGCCTGATG	GAA W TGG L CTG G GGT 8000 S AGC A GGCG * TCC
S T D M T K W TCTACCEATATGACCAMETGCC 4450 S T V T A P Q TCAMCCETATGCCETTCCMC A W G R R G L CCEMACETACCCETTCCMC 4700 P L V F G P Q TTCCTGCTTTTGCCCCCAMI P A K L G L G TTTCCCAMETTACCTCTGCAMI C I V A V M E E ATGCTGCCCTAMEGAMAMA 5050	CONCONTRACTOR ACCORTANT 4350 Q I F V E Y V L MARCENTERTENENTETERTE E L K L S K A V INCENTION TO THE ACCORTON 4600 V A C I A L A L TROCCESCIENT CONCECTOR I L I G V A A V TO THE ACCORT OF CONCECTOR M I A D G I P V NOANTROCCESTROGENEOCCOCTANT 4950 R K I R R E K K CONAMMICCOCCOCCECTANTANA	K W K V I W L L ANAMCAMACTGATCTGGCTGGC 4500 A I Q G F T L F SCGATCCMGGCCTTTMCGCTGTT 1 I A T L G V Y ATTATCGCCMCGCTGGGGGGGGGGGGGGGGGGGGGGGGGG	AGCTGANGANCTGTTCGGCGAG 4400 C F A N I F L GTGCTTCGCCANCATTTTCCTC * E A G A L V G TGANGCTGGTGGCGCCGGTGGGT 4650 Q H A S N E Y TCANCATGCCGGTANGGANTAN 50 G A A D G I K TGCGCCGCCGCGGTGGTATTAN * T F A A L D I CACCTTGCCCGCGCGGTGGATATC 5000	CACATCHOCCAMERICANGACAAAGACAA Y V V R I G I D Q TATGTGCTACGTATTGGTATCGACCAG * 4550 T L L W G W L S D ACCCTOCTGCGCCTGCTGCCAC * 4 I Y L A S L F A L ACCCTACTGGCCTGCCTGCTGTGC * 4 G T F A Y L I G D AGCCCCCTTGCTTACTGATTGGTGAC 4900 A A I G C I C L M ***********************************	GAA W TGG L CTG G GGT 800 S AGC * A GGCG * TCC *
S T D N T K W TCTACCENTATENCIMINACE 4450 S T V Y A P Q TCANCESTATACCECTTOCASE COMMERCIAL CONSTRUCTION P L V F C P Q THCCTGCTCTTTGCCCCCCCAN P A K L C L C TTTCCCAMETTAGCTCTOCAN I V A V M E ATASTCCCCCTANTCCAMENAN 5050	CONCONTINUOUS ACCALCENT 4350 Q I F V E Y V L MARCHTTGTTGAGTATGTGCTG E L K L S K A V IMACTGAAACTCYTAACCGGTG 4600 V A C I A L A L TTGGCCTTGCATCGCGCTGCGCTG L L I G V A A V TTGTTGATTGCTGTGGTGGCGCGCGGTA 4950 R K I R R E K K SCHAMANTCCCCGCGGGGGAGAAAAA	K W K V I W L L AMANACANATGATCTGGCTGGT 4500 A I Q G F T L F SCGATTCAGGCCTTTACGCGTTTA I I A T L G V Y ATTATCGCCACGCTGGGTGGGA 47 G F V P K K A I SCGTTGTACCTANANAGGAT 4550 F G L T G W A G TTCGGCCTTACCGGGTGGGCAGG ATTCCGCCACTGACGTGGGCAGA S100	ACCTGANGANCTETTCCCCGAN 4400 C F A N I F L GTGCTTCGCCANCATTTTCCTC * E A G A L V G TGANGCTGGTCGCCTGGTCGGT * 4650 Q H A S N E Y TCANCATGCCGATGGTATTAN 50 G A A D G I K TGGCGCTGCCGATGGTATTAN 50 T F A A L D I CACCTTCGCCCGCCGGTATATC 5000	CACATCHACCAMERICANCANAGACAA Y V V R I G I D Q TANGTGCTANGTATGGTATGGACCAG 4550 T L L W G W L S D ACCOTOTOGOGCCTGCCTCTCAC Y L A S L F A L ACCTATCGCCTTCTCCTTGCGTG 4 G T F A Y L I G D GCCCCCCTTCCTTCCTGCTGCCTGCCTGATG 4900 A A I G C I C L M CCCCCCGATTGGTGTATCTGCCTGATG * *	GAA W TGG CTG G GGT 8000 S AGCC * A CTCC *
S T D M T K W TCTACCENTATERCEAMSTREE 4450 S T V T A P Q TCAACCETATACCECTTCAAC 4450 P L V F C P Q TTOCTGETCTTTCCCCCCAAT P A K L C L C TTTCCCAACTTAGTCTCGCAAT S I V A V M E E ATACTGCCTAATCAAL 5050	COCOCONTRACTOCACCACACTAC 4350 0 I F V E Y V L MARCETTETTETTENENTETECTEC E L K L S K A V INACTORANCTOCTANECCETTE 4600 V A C I A L A L TENECCTECATOCCCTECTETTE N I A D G T P V MIA TOCCATOCCCCETA 4950 R K I R R E K K COCAMANTCCCCCCCATAMANAC	K W K V I W L L NANACANAGGARCTGGCTGGC 4500 A I Q G F T L F SCGATCKAGGGCTTTACGCGTGT 1 I A T L G V Y MTTATGGCCAGGCGGGGGGGGGGGGGGGGGGGGGGGGGGG	AGCTGANGANCTGTTCGGCGAG 4400 C F A N I F L GTGCTTCGCCANCATTTTCCTC E A G A L V G TGANGCTGGTGCCTGGTGGGT 4650 Q H A S N E Y TCANCATGCCGTGATATAN G A A D G I K TGGCGCTGCCGCGTGGATATAN T F A A L D I CACCTTGCCGCGCCGCGTGATATC 5000 AMCGTANCTGGTGACTTTTGCC 2000	CACATCHOCCAMERICANCANGUARCA Y V V R I G I D Q TATGTOGTACGTATTGGTATCGACCAC # 4550 T L L W G W L S D ACGOTOCTGTOGGCTGGCTGCTGTGC I Y L A S L F A L I Y L A S L F A L I Y L A S L F A L G T F A Y L I G D MGCCCCGATTGGTGTGTGTGTGGTGAC 4900 A A I G C I C L W CCCCCGATGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGT	GAA W TTGG L CTG G GGT 8000 S AGCG X AGCG X CTCC X GGT
S T D N T K W TCTACCENTATCACCENTATCACTACTACT 4450 S T V Y A P Q TCANCESTATACCESTTOCACC COMMENSATIACCESTTOCACC 4700 P L V F G P Q TTOCTESTCTTOCCCCCCA F A K L G L G TTTCCCACESTATACCECTOC S050 GTCACTTCCCCCSTATCCANCAC	CONCONTRACTOR ACCALCE 4350 Q I F V E Y V L MARCHINGTANGTANGTANGTOCTO E L K L S K A V MARCHINGTONTONOCTON V A C I A L A L TOUTONATOCTOCOCTOCOCTO L L I G V A A V TOUTONATOCTOCOCTOCOCTOCTOT M I A D G T P V MONTRACTOCTOCOCCOCCOCTANA 4950 R K I R R E K K COCANANTCCCOCCOCCANCANANA COMMETTOCTOCOCCOCCANCANANA COMMETTOCTOCOCCOCCANCANANA COMMETTOCTOCOCCOCCANCANANA COMMETTOCTOCOCCOCCANCANANA	K W K V I W L L ANAMCANASTGATCTGGCTGCT 4500 A I Q G F T L F SCGATTCHOGGCTTTMCGCTGTT I I A T L G V Y ATTATCGCCACGCTGGGTGGGTA 4550 F G L T G W A G FTCGCCTTACCTANAAAACCAT 4550 F G L T G W A G TTCGCCTTGCTGGCTGGGCATA 5100	AGCTGANGANCTGTTCGGCGAG 4400 C F A N I F L GTGCTTCGCCANCATTTTCCTC * E A G A L V G TGANGCTGGTGGCTGGGTGGGT GANGCTGGCGCCGGCTGGATGGT G A A D G I K T F A A L D I CACCTTCGCCGCGCGCATATAT 5000 ANCGTANCTGGCGCGCGCGCTGATATTCC 5000	CACATCHOCCAMERICANCANGLANCA Y V V R I G I D Q TATGTGGTACGTATTGGTATCGACCAG * 4550 T L L W G W L S D ACCOCCATGGGGGCTGGCTGCTGCTGC I Y L A S L F A L I Y L A S L F A L C T F A Y L I G D C T F A Y L I G C A A I G C I C L M CCCCCGATTGGTTGGTTGTATCTGCTGATG * 5150 MACCCCATCGGGCCTATTGTTTATTAT * 5150	GAA W TGG CTG G GGGT S S GGGT X GGGT C C C C C C C C C C C C C C C C C
S T D N T K W TCTACCEATATCAACAAASTOCC 4450 S T V Y A P Q TCAACCETATACCATTCCAGO A W G R R G L GCGAACCETCCCCCCCCCCCCCCC 4700 F L V F G P Q TTCCTGCTCTTTGGCCCCCCAAT P A K L G L G TTTCCCAACTTAGCTCTGGAA S050 GTGACTTCCACCETAGTGAACAA	CONCONTRACTOR ACCALCE 4350 Q I F V E Y V L MEATCHTHETTERETERETERETERE E L K L S K A V MACTERAACTOCTARACCOSTO 4600 V A C I A L A L NEDECTECATOCCOCCOCCACT I L I G V A A V HENTERTECTOCATOCCCCCCCTA 4950 R K I R R E K K COLAMATCCCCCCCCCATCAAAAAA 5200	K W K V I W L L ANANACANAGGATCTGGCTGCT 4500 A I Q G F T L F SCGATTCHOGGCTTTACGCGTGTT I I A T L G V Y MTTATCGCCAGCCCGGGTGGGTTA 477 G F V P K K A I SCGTTTGTACCGAGTGGGTAA F G L T G W A G F C L T G W A G F C L T G W A G F C C L T C V A * ATTCHCCCAGTGCCGGCTAGGCTAA 5100 CCCCGGCTGGTTAGGTTTAGCGC	AGCTGANGANCTGTTCGGCGAG 4400 C F A N I F L GTGCTTCGCCAMCATTTTCCTC E A G A L V G TGANGCTGGTGCGCTGGTCGGT 4650 Q H A S N E Y TCANCATGCCMGTANGGANTAJ 50 G A A D G I K TGGCGCTGCCGGCTGGATATAG 500 ANCGTANCTGGTGACTTTTCCC * * * 5250	CACATCACCOALGACAACAAAGACAA Y V V R I G I D Q TATGTGCTACGTATTGCTATCGACCAC 4550 T L L W G W L S D ACCTOCTGGGGCGCGGCCTGCTGC I Y L A S L F A L I Y L A S L F A L G T F A Y L I G D GGCACCTTRCCTTACCTGATGGCGC A A I G C I C L W COCCACTACGCTGTATCTGCTGATGG T COGCACGATGGCCGGGCTTTTTTATTAT * 5150	GAA W TGG CTG G GGT S AGC C C C C C C C C C C C C C
S T D M T K W TCTACCEATATGACCAMETGOC 4450 S T V T A P Q CAMCCETATGACCAMETGOC 4450 S T V T A P Q A W G R R G L CCCMACGTAGCOTTOCAGO 4700 P L V P G P Q TTOCTGGTCTTTOCCCCCAMAT P A K L G L G 1TTOCCAMETTAGCTCOGAM S050 * AMCGACGCCTATGCCAMAGE * AMCGGACGCCATTCACCOGCOM	CONCONTRACTOR 4350 0 I F V E Y V L MARCENTERTRAGENERGE E L K L S K A V INACTORANCECTERTRAGENERGE 4600 V A C I A L A L TREECENSERGENERGENERGENER L L I G V A A V TESTERTRATEGENERGENERGENER M I A D G I P V NORTHOCONTROCHEGENERGENER 4950 R K I R R E K K SCHAMMECCECCOCCACEALANA SCHAMETERCECCOCCACEALANA SCHAMETERCECCOCCACEALANA SCHAMETERCECCOCCACEALANA SCHAMETERCECCOCCACEALANA SCHAMETERCECCOCCACEALANA SCHAMETERCECCOCCACEALANA SCHAMETERCECCOCCACEALANA SCHAMETERCECCOCCACEALANA SCHAMETERCECCOCCACEALANA SCHAMETERCECCOCCACEALANA SCHAMETERCECOCCACEALA	K W K V I W L L ANAMCANACTGATCTGGCTGGT 4500 A I Q G F T L F SCGATTCNGGGCTTTMCGCTGTT I I A T L G V Y ATTATCGCCACGCTGGGTGTGTA 47. G F V P K K A I SCTTTGTACCTANAANGCGAT 450 F G L T G W A G FTCGGCCTTACCGGCGGGGGGGGGGGGGGGGGGGGGGGGG	AGCTGAAGAACTGTTCGGCGAG 4400 C F A N I F L GTGCTTCGCCAACATTTTCCTC * E A G A L V G G A A C A L V G Q H A S N E Y TCAACATGCCGGTAGGATAGAATAT 50 G A A D G I K TGCGCCGCCGCGGCGATATAAJ * T F A A L D I CACCTCGCCGCGCGCGGCATATAC 5000 AACGTAACTGGTGACCTTTTGCC * GGGGATCACTGGCGAGGAAAAGG * 5250	CACATCHOCCHACHOGACHAGACHA Y V V R I G I D Q TATGTOGTACGTATTGGTATCGACCAG * 4550 T L L W G W L S D ACCOTOCTOGOGOCTGGCTCTCTAC I Y L A S L F A L I Y L A S L F A L G T F A Y L I G D ACCOTOCATTGCTTACCTGATTGGTGAC 4900 A A I G C I C L M * 5150 COOCATCACCCCCCCCTTTTTATTAT * 5150	GAA W TGG CTG G GGT S S CTG C C G G G CTG C C C C C C C C C C C C C
S T D N T K W TCTACCEATATCACCAMPACA 4450 S T V Y A P Q TCANCCETATACCCTTCCACC A W G R R G L CCANCECTCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	BCCTCOCFTRCOCCACCALCOC 4350 Q I F V E Y V L MARCHTRETTERSTANGTATEGECTE E L K L S K A V MARCHTRETTERSTANGTATEGECTE 4600 V A C I A L A L CTOCCTCATCOCCTOCATCOCCTOCATEGECTECCTETTER M I A D G T P V M I A D G T P V MORTTEGENTECTESCECCECTETT 4950 R K I R R E K K SCEAMACTCCTCCCCTCATEGECACCCCCTATAGEC SCANACTCCCCCCCCCTCATAGECACCCCCTATAGEC SCANACTCTCCCCCCCCCTCATAGECACCCCCTCATAGECACCCCCCTCATAGECACCCCCCCTCATAGECACCCCCCTCATTGET	K W K V I W L L ANANACANATATAGGETEGGETAGE 4500 A I Q G F T L F SCGATTCAGGECTTTAGGEGETTA I I A T L G V Y ATTATCGCCAGGECTGGGEGGEGGE 47 G F V P K K A I SCTTTGTACCTANANAGGAT 4550 F G L T G W A G TTCGCCTTACTGGGEGGGCAGE I Q Q L T V A * ATTCCCCAGGETTGGTTGGTTTAGGEGGAGE S100 CCCCCGACCGGCTTGGGCCGGTGGG S3	AGCTGAAGAACTGTTCGGCGAG 4400 C F A N I F L GTGCTTCGCCAACATTTTCCTC * E A G A L V G TGAAGCTGGTCGCCTGGTCGGT * 4650 Q H A S N E Y TCAACATGCCCGTAGGTAATAG 50 G A A D G I K TGGCGCTGCCGATGGTATTAA 50 T F A A L D I CACCTTCGCCCGCGCTGGATATC 5000 AACGTAACTGGTCACTGTTGCC * 5250 GGGGATCACTGGCCGGCAAAG * 5250	CACATCHOCOLAGE ACANGE ACAN T V V R I G I D Q TATGTOGTACGTATGGTATGGTATGGACCAG 4550 T L L W G W L S D ACCOTOTOGOGCTOCCTOTCAC I Y L A S L F A L I Y L A S L F A L I Y L A S L F A L G T F A Y L I G D GC T F A Y L I G D GC T F A Y L I G D GC C I C L M COCCATCATACTACTACTACTACTACTAC 4900 A A I G C I C L M COCCATACACCOCCTTATTATAT * 5150 MACCOCCATCACATATAAACCOCTCATCG *	GAA * W TTGG CTG G G G G G G G G G C G G C G G C G G C G G C G G C G

generate a 199-amino-acid polypeptide with a molecular weight of 21,787. Upstream of this start site is a polypurine stretch, GGGG, which could serve as a ribosome-binding sequence. Transposon insertion 33 in uhpC is located between nucleotides 2462 (*Eco*RV-3) and 2690 (*HpaI*-2), consistent with the identification of this reading frame as uhpC. The location of the actual start site must await analysis of the amino-terminal polypeptide sequence.

The reading frame for the uhpT gene extends between nucleotides 3718 and 5106 and encodes a 463-amino-acid polypeptide of M_r 50,615. All of the transposon insertions that specifically inactivate uhpT are located within this region. A Shine-Dalgarno sequence (AGGAG) is centered 8 nucleotides upstream from the coding region. The termination codon is followed by a typical Rho-independent transcription terminator structure, namely, a 7-base-pair GCcontaining stem with a 5-nucleotide loop ($\Delta G = -21.9$ kcal/ mol [ca. -91.6 kJ/mol]) (22), and followed by six contiguous T residues (21).

Identification of potential transcriptional regulatory signals. There is a potential promoter sequence located between nucleotides 99 and 137, just upstream from the uhpA coding region. This sequence contains a -35 TTGAAC and a -10 TAAGGT region separated by 17 nucleotides; the invariant T residue in the -10 region is located 7 nucleotides upstream from a potential transcription initiation site (TAT) (21). The poor match of this putative promoter sequence with the

canonical sequence is consistent with the relatively low level of expression of the *uhp* regulatory genes (unpublished data). This promoter overlaps the *ilvBN* terminator such that the run of T residues in the terminator would be the first nucleotides in the *uhpA* transcript.

Since transposon insertions in uhpA have a polar effect on expression of uhpB (32), it is likely that both genes are transcribed from the same promoter into a polycistronic message. No promoter sequences were detected in front of uhpB, but a weak promoter might have been missed. The poor expression of uhpC in maxicells makes it uncertain whether this gene is part of the same operon as the other regulatory genes. Transposon insertion 28, lying between uhpB and uhpC, does not confer a Uhp⁻ phenotype, suggesting that the insertion does not eliminate expression of uhpC. However, these results do not preclude a reduction in synthesis of this protein. Potential weak promoter sequences can be found upstream of uhpC (for example, between nucleotides 2542 and 2556), but determination of their significance will require transcript mapping.

A potential promoter region is seen in front of uhpT. A transcription initiation sequence CAC (nucleotides 3689 to 3691) is located 17 nucleotides before the Shine-Dalgarno sequence. This start site is preceded by a typical -10 sequence TACAATG 7 nucleotides upstream. There is, however, no sequence resembling the consensus -35 sequence at the expected location. Instead, the symmetrical



FIG. 3. Hydropathy profiles of the four Uhp polypeptides. The hydropathy parameters of Kyte and Doolittle (15) were averaged over a moving window of seven amino acids. The four polypeptides are (A) UhpA, (B) UhpB, (C) UhpC, and (T) UhpT. Values above the zero axis represent hydrophobic segments.

sequence TCAGGCCTGA is centered at the -35 position. Lying upstream from this region are numerous sequences of dyad symmetry and direct repeats, which may represent sites for binding of regulatory proteins. Further upstream, between nucleotides 3573 and 3593, at the area which lies at -97 to -117 relative to the transcript start site, is a typical catabolite activator protein-binding site. This sequence matches the consensus catabolite activator protein-binding site at 11 of 14 positions. Definition of the role of these sequences in promoter function is in progress. The absence of a canonical promoter sequence was expected for a promoter whose expression is dependent on a positive activator protein (20).

Codon usage. Grosjean and Fiers (11) noted a substantial difference in the distribution of codons between strongly and weakly expressed genes in E. *coli*. The preferences were ascribed to the optimization of the energy of the codon-anticodon interactions and to the amounts of the various isoaccepting tRNA species. So-called "modulator" codons are used rarely in highly expressed genes, whereas most of the codons are present in weakly expressed genes.

There are differences in the codon usage patterns of the four *uhp* genes which appear to be consistent with the expected differences in their levels of expression. The *uhpA*, *uhpB*, and *uhpC* genes contain fewer of the preferred codons, defined by Bennetzen and Hall (4), than does *uhpT*, i.e., 53, 48, and 46% versus 63%, respectively. Similarly, the three regulatory genes contain more modulator codons (6.1, 11.3, and 12.4%, respectively) than does *uhpT* (1.9%). Thus, the codon usage of *uhpT* is consistent with that of a relatively strongly expressed gene, whereas those of *uhpB* and *uhpC*

are characteristic of weakly expressed genes and that of uhpA seems intermediate.

Another parameter of codon bias, termed P2', was defined by Sharp and Li (24) and represents the preference for codons of intermediate base-pairing energy, i.e., those that contain both A-U and G-C pairings as opposed to those having only A-U or only G-C pairs. This P2' parameter was proposed to be higher for more highly expressed genes. However, the P2' values for the *uhp* genes were roughly the same, in contrast to the differences in codon bias seen with the parameters described above.

Characteristics of the predicted Uhp polypeptides. The properties of the Uhp polypeptides predicted from the nucleotide sequence compare favorably with the size and locations of the actual Uhp proteins observed in maxicells. UhpA is a soluble protein, whereas UhpB and UhpT are membrane associated. The predicted UhpA polypeptide has a content of charged amino acids typical of soluble proteins (Asp+Glu+Arg+Lys = 21 mol%), whereas the other three Uhp proteins are less polar (18.3, 12.8, and 14.5 mol% for UhpB, UhpC, and UhpT, respectively). UhpA has a net negative charge, and the other three proteins bear net positive charges.

Kyte and Doolittle (15) have developed a method for calculating the hydropathic character of a protein and for displaying the distribution of polar and nonpolar residues along its primary sequence. The average hydropathy values for UhpA and UhpB are 0.18 and 0.19, respectively, which are in the range of soluble proteins. UhpC and UhpT had more nonpolar character, with net hydropathy values of 0.59 and 0.56, respectively.

Protein	Starting residue	Sequence"	Reference
UhpA	153	MAVKEIAAELGL SPK TVHVHRA	
cİ	31	LSOESVADKMGMGQSGVGALFN	19
Cro	17	F GÕTKTAK D L G V YÕ S A I NLA I H	19
Cro 434	19	M TOTELATK AGV KOQ S IQLIEN	10
TrpR	66	M SQRELKNE LGAGIAT I TRGSN	
DeoR	22	L HLKDAAA L LGV SEM T I RRDL N	30
CAP ^b	167	I L R Q E I G Q I V G C S R E T V G R I L K	27
AsnC	23	T AYAELAKQ FGV SPG T I HVRV E	14
Fnr	195	M T R G D I G N Y L G L T V E T L S R L L G	27
OmpR	88	E V D R I V G L E I G A D D Y I P K P F N P	
PhoB	199	I RRLRKAL E PGG HDR M VQT V R G	17

TABLE 1. Comparison of a portion of UhpA with DNA-binding regions of other regulatory proteins

^a Amino acid residues are in standard one-letter code. Residues in UhpA with a double underline are those in which the same residue is present in at least three of the other regulatory proteins; those with a single underline are replaced by homologous amino acids in at least seven of the other proteins.

^b CAP, Catabolite activator protein.

Regions of high hydropathic character that are at least 20 amino acids in length might traverse the cytoplasmic membrane (8). The hydropathy profiles of the four Uhp proteins (Fig. 3) lead to definite predictions about the interaction of the proteins with the membrane. The UhpA protein has a hydropathy profile typical of a soluble protein, lacking long nonpolar stretches that might span the membrane. In contrast, UhpB, which is associated with the membrane fraction despite its overall polar character, consists of two distinct domains. The amino half of the protein is strongly nonpolar and has as many as nine potential membrane-spanning stretches. The carboxyl half of the protein is very polar and displays no regions of high hydrophobic character. Thus, it is likely that UhpB is embedded in the membrane by multiple traverses within its amino half, while the carboxyl half resides exclusively in either the periplasmic space or the cytoplasm.

The UhpC protein possesses four to six potential membrane-spanning segments. Based on this hydropathy profile and its overall nonpolar character, it is probable that UhpC is membrane associated, although this has not been demonstrated directly. The UhpT transport protein also contains numerous potential membrane-spanning regions distributed along its entire length, although it is more polar than other cytoplasmic membrane transport proteins such as the lactose permease (6) or the integral membrane components of periplasmic binding protein-dependent transport systems (2).

Homology of the Uhp proteins. Since the uhpA product appears to activate uhpT transcription (26, 32), it is likely that it binds to specific DNA sequences and hence might display sequence homologies with other DNA-binding proteins. DNA-binding proteins of known crystal structure, such as catabolite activator protein, Cro, cI, and TrpR, interact with their target sites through specific hydrogen bonds made by amino acid side chains present in a characteristic helix-turn-helix motif (19). Many other regulatory proteins possess sequence homologies to this bihelical region (10). Sequences homologous to this conserved 22amino-acid region are found near the C terminus of UhpA (amino acid residues 153 to 174) and are compared with those of several other regulatory proteins in Table 1. Note that there is no obvious difference within this 22-amino-acid region between proteins that function as repressors and those that activate gene transcription. Of the regulatory proteins examined, UhpA shares the most amino acids in

this region with AsnC, an activator of the asnA (asparaginase) gene (14).

The UhpT protein shares about 30% amino acid identity along its entire length with the GlpT glycerol 3-phosphate transport protein (W. Boos, personal communication). Both of these transporters appear to act by a phosphate-antiport mechanism (1). No substantial homology was seen between UhpT and any of the proteins present in the National Biomedical Research Foundation Protein Sequence Database.

A very striking homology exists between the regulatory protein UhpC and the transporter UhpT. In UhpC, this region of homology begins at amino acid residue 39 and continues to the end of the protein. This stretch matches UhpT between residues 160 and 345 and is characterized by 33% amino acid identity over the 185-amino-acid overlap. If one includes conservative amino acid substitutions, these sections of the two proteins are 66% homologous. Not only are the amino acid sequences conserved between these two proteins, but so too are their possible transmembrane orien-



FIG. 4. Comparison of the hydropathy profiles of UhpC and the homologous portion of UhpT. The Kyte-Doolittle hydropathy profile of UhpC (solid line; coordinates on bottom ordinate) is compared with that of the region of UhpT with conserved amino acid sequence (dotted line; coordinates on top ordinate).

tations. The Kyte-Doolittle hydropathy distributions for UhpC and the corresponding portion of UhpT are portrayed in Fig. 4. The close correspondence within the homologous regions suggests that the two proteins have very similar conformations and orientations, although local differences exist.

The nucleotide sequence encoding the 185 conserved amino acids (aligned by five insertions of an amino acid in one or the other protein) exhibits 45.7% identity. The conservation of nucleotide sequence tends to be clustered and separated by stretches of unrelated sequence. Of the 60 identical amino acids present in analogous sites in the two proteins, more than half use the same codon. This suggests that the conserved portions of both proteins evolved from a common sequence.

The most likely mechanism for such identity is that the UhpC regulatory protein evolved by partial gene duplication from the UhpT transporter, whose synthesis it regulates. These two proteins could share similar structural requirements if one assumes that UhpC is the regulatory protein that binds the inducer and acts at the first step of the signaling process (32). Thus, both UhpC and UhpT proteins would have to span the membrane and possess a substratebinding site accessible to the exterior. One factor that would prevent UhpT itself from serving this regulatory role is that its substrate specificity for transport function is much less stringent than the specificity for the inducer (reviewed in reference 32), arguing that the same binding site cannot be employed for both transport and regulation. To our knowledge, this is the first example where a regulatory protein might have evolved from the gene it regulates.

Implications for mechanism of regulation. One of the interesting features of the uhp system is its regulation by external inducer independent of UhpT transport function. The simplest model for exogenous induction involves the interaction of a transmembrane regulatory protein(s) with the inducer on the periplasmic face of the cytoplasmic membrane, which causes a transmembrane signaling event that activates the uhpA product to allow transcription of uhpT.

Properties of the predicted *uhp* polypeptides provide some support for this model. The UhpA activator protein contains a region with substantial homology to the areas of DNAbinding regulatory proteins that are responsible for the sequence-specific interactions between the protein and its DNA target (19). It remains to be determined where the UhpA protein binds in the *uhpT* promoter region and whether its activity is regulated by some covalent modification or by its release from the membrane.

Both UhpB and UhpC seem to be associated with the cytoplasmic membrane, although this has been directly shown only for UhpB. Both possess multiple stretches that have high hydropathic character and are at least 20 amino acids long, sufficient to span the membrane bilayer as an α helix. The amino-terminal half of UhpB could be almost totally embedded in the membrane, with the carboxy-terminal half folded on one side of the membrane. Predictions of the distribution of UhpB across the membrane suggest that few residues are exposed on the side of the membrane opposite the carboxyl half. These predictions suggest that UhpB has little transmembrane character, but rather that half is buried in the membrane and the other half is completely on one side of the membrane.

Hydropathy plots of UhpC are less amenable to prediction of transmembrane distribution, although several potential membrane-spanning regions are apparent. Substantial portions of this protein are expected to be exposed on either side of the membrane.

Genetic studies led to the proposal that UhpC might serve as receptor for the inducer. Mutations in uhpA or uhpBresult in a Uhp⁻ phenotype that is rarely, if ever, reversed by second-site mutations. In contrast, the Uhp⁻ phenotype incurred by mutations in uhpC are suppressed by frequent events within another *uhp* gene or genes; the site of these mutations is unknown (13, 32). In most of these revertants, UhpT expression is no longer induced by the presence of glucose 6-phosphate. The strong homology of UhpC with UhpT is consistent with the presence on UhpC of a sugarphosphate-binding site. Thus, our current model for regulation can account for the necessity for all three regulatory proteins by assuming that UhpB modifies UhpA in some way to convert it to an active form. The constitutive expression seen with the elevated copy number of uhpA is independent of UhpB and may reflect the increased equilibrium concentration of the activated conformer of UhpA. Direct evidence for several features of this model are being sought, and alternatives cannot be excluded.

ACKNOWLEDGMENTS

We gratefully acknowledge discussions with Lucy Weston concerning the Uhp polypeptides and their genetics, with Sarah French on manuscript presentation, and with Mitch Smith and Michael Lundrigan on DNA sequencing and computer analyses.

This work was supported by research grant PCM-8215915 from the National Science Foundation and by Public Health Service grant GM38681 from the National Institute of General Medical Sciences.

LITERATURE CITED

- Ambudkar, S. V., T. J. Larson, and P. C. Maloney. 1986. Reconstitution of sugar phosphate transport systems of *Escherichia coli*. J. Biol. Chem. 261:9083–9086.
- Ames, G. F.-L. 1986. Bacterial periplasmic transport systems: structure, mechanism, and evolution. Annu. Rev. Biochem. 55:397-425.
- 3. Barr, P. J., R. M. Thayer, P. Laybourn, R. C. Najarian, F. Seela, and D. R. Tolan. 1986. 7-Deaza-2'-deoxyguanosine-5'-triphosphate: enhanced resolution in M13 dideoxy sequencing. Biotechniques 4:428-432.
- Bennetzen, J. L., and B. D. Hall. 1982. Codon selection in yeast. J. Biol. Chem. 257:3026–3031.
- Biggin, M. D., T. J. Gibson, and G. F. Hong. 1983. Buffergradient gels and ³⁵S label as an aid to rapid DNA sequence determination. Proc. Natl. Acad. Sci. USA 80:3963-3965.
- 6. Buchel, D. E., B. Gronenborn, and B. Muller-Hill. 1980. Sequence of the lactose permease gene. Nature (London) 283: 541-545.
- 7. Dale, R. M. K., B. A. McClure, and J. P. Houchins. 1985. A rapid single-stranded cloning strategy for producing a sequential series of overlapping clones for use in DNA sequencing: application to sequencing the corn mitochondrial 18S rDNA. Plasmid 13:31–40.
- Engelman, D. M., T. A. Steitz, and A. Goldman. 1986. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Annu. Rev. Biophys. Biophys. Chem. 15: 321-353.
- 9. Friden, P., J. Donegan, J. Mullen, P. Tsui, and M. Freundlich. 1985. The *ilvB* locus of *Escherichia coli* K-12 is an operon encoding both subunits of acetohydroxyacid synthase I. Nucleic Acids Res.13:3979–3993.
- Gicquel-Sanzey, B., and P. Cossart. 1982. Homologies between different prokaryotic DNA-binding regulatory proteins and between their sites of action. EMBO J. 1:591–595.
- 11. Grosjean, H., and W. Fiers. 1982. Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed

genes. Gene 18:199-209.

- Kadner, R. J. 1973. Genetic control of the transport of hexose phosphates in *Escherichia coli*: mapping of the *uhp* locus. J. Bacteriol. 116:764-770.
- Kadner, R. J., and D. M. Shattuck-Eidens. 1983. Genetic control of the hexose phosphate transport system of *Escherichia coli*: mapping of deletion and insertion mutations in the *uhp* region. J. Bacteriol. 155:1052–1061.
- Kolling, R., and H. Lother. 1985. AsnC: an autogenously regulated activator of asparaginase synthetase A transcription in *Escherichia coli*. J. Bacteriol. 164:310–315.
- 15. Kyte, J., and R. F. Doolittle. 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157:105-132.
- Lipman, D. J., and W. R. Pearson. 1985. Rapid and sensitive protein similarity searches. Science 227:1435–1441.
- Makino, K., H. Shinagawa, M. Amemura, and A. Nakata. 1986. Nucleotide sequence of the *phoB* gene, the positive regulatory gene for the phosphate regulon of *Escherichia coli* K-12. J. Mol. Biol. 190:37-44.
- Messing, J. 1983. New M13 vectors for cloning. Methods Enzymol. 101:20-78.
- Pabo, C. O., and R. T. Sauer. 1984. Protein-DNA recognition. Annu. Rev. Biochem. 53:293-321.
- Raibaud, O., and M. Schwartz. 1984. Positive control of transcription initiation in bacteria. Annu. Rev. Genet. 18:173–206.
- 21. Rosenberg, M., and D. Court. 1979. Regulatory sequences involved in the promotion and termination of RNA transcription. Annu. Rev. Genet. 13:319–353.
- Salser, W. 1977. Globin mRNA sequences: analysis of base pairing and evolutionary implication. Cold Spring Harbor. Symp. Quant. Biol. 42:985-1002.
- Sanger, F., S. Nicklen, and A. R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA 74:5463-5467.

- Sharp, P. M., and W. H. Li. 1986. Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. Nucleic Acids Res. 14:7737-7749.
- Shattuck-Eidens, D. M., and R. J. Kadner. 1981. Exogenous induction of the *Escherichia coli* hexose phosphate transport system defined by *uhp-lac* operon fusions. J. Bacteriol. 148: 203-209.
- Shattuck-Eidens, D. M., and R. J. Kadner. 1983. Molecular cloning of the *uhp* region and evidence for a positive activator for expression of the hexose phosphate transport system of *Escherichia coli*. J. Bacteriol. 155:1062–1070.
- 27. Shaw, D. J., D. W. Rice, and J. R. Guest. 1983. Homology between CAP and Fnr, a regulator of anaerobic respiration in *Escherichia coli*. J. Mol. Biol. 166:241-247.
- Shine, J., and L. Dalgarno. 1974. The 3' terminal sequences of Escherichia coli 16S rRNA: complementarity to nonsense triplets and ribosome binding sites. Proc. Natl. Acad. Sci. USA 71:1342-1346.
- Staden, R. 1980. A new computer method for the storage and manipulation of DNA gel reading data. Nucleic Acids Res. 8:3673-3694.
- Valentin-Hansen, P., P. Hojrup, and S. Short. 1985. The primary structure of the DeoR repressor from *Escherichia coli* K-12. Nucleic Acids Res. 13:5926–5936.
- Wek, R. C., C. A. Hauser, and G. W. Hatfield. 1985. The nucleotide sequence of the *ilvBN* operon of *Escherichia coli*: sequence homologies of the acetohydroxy acid synthase isozymes. Nucleic Acids Res. 13:3995–4010.
- Weston, L. A., and R. J. Kadner. 1987. Identification of Uhp polypeptides and evidence for their role in exogenous induction of the sugar phosphate transport system of *Escherichia coli* K-12. J. Bacteriol. 169:3546-3555.
- Winkler, H. H. 1970. Compartmentation in the induction of the hexose-phosphate transport system in *Escherichia coli*. J. Bacteriol. 101:470–475.