

## NOTES

# Tandem Translation Starts in the *cheA* Locus of *Escherichia coli*

ERIC C. KOFOID AND JOHN S. PARKINSON\*

Biology Department, University of Utah, Salt Lake City, Utah 84112

Received 9 October 1990/Accepted 6 January 1991

The *cheA* locus of *Escherichia coli* encodes two protein products, CheA<sub>L</sub> and CheA<sub>S</sub>. The nucleotide sequences of the wild-type *cheA* locus and of two nonsense alleles confirmed that both proteins are translated in the same reading frame from different start points. These start sites were located on the coding sequence by direct determination of the amino-terminal sequences of the two CheA proteins. Both starts are flanked by inverted repeats that may play a role in regulating the relative expression rates of the CheA proteins through alternative mRNA secondary structures.

The *cheA* locus of *Escherichia coli*, which is required for chemotactic behavior, encodes two cytoplasmic proteins, CheA<sub>L</sub> and CheA<sub>S</sub>, of apparent molecular weights 78,000 and 69,000, respectively (7, 8). Nonsense mutations throughout most of the *cheA* coding region truncate both proteins, demonstrating that CheA<sub>L</sub> and CheA<sub>S</sub> are made in the same reading frame. However, several nonsense mutations at the promoter-proximal end truncate only CheA<sub>L</sub>, indicating that they lie outside of the coding sequence for CheA<sub>S</sub>. These observations led to the suggestion (8) (Fig. 1) that CheA<sub>L</sub> and CheA<sub>S</sub> were made by initiating translation of the *cheA* mRNA at two different in-frame start sites, which we denote as start(L) and start(S).

To test the two-start model of *cheA* expression, we determined the nucleotide sequence of the wild-type *cheA* locus and the N-terminal amino acid sequences of its two protein products. Our findings not only support the model, but also imply that competitive interactions between start(L) and start(S) may be important in regulating the relative expression levels of the two CheA proteins.

**Nucleotide sequence of the *cheA* locus.** The *cheA* locus lies in the middle of an operon containing several other genes, *motA* and *motB* upstream, and *cheW* downstream. A restriction fragment spanning the entire *cheA* operon was obtained by *XmaI-XbaI* digestion of  $\lambda$ che22 DNA (5) and cloned into the corresponding sites of plasmid pUC118 (11), yielding pEK46 (Fig. 1). Single- or double-stranded DNA from pEK46 or one of its derivatives (not shown) was used as the template for dideoxy sequencing reactions (6). Synthetic oligonucleotides complementary to the cloned insert were used as primers. Sequence was determined on both strands from the *AccI* site in *motB* through the *XbaI* site at the end of the insert (Fig. 1). The sequence of the *cheA* coding region and pertinent flanking features is shown in Fig. 2.

An open reading frame of 654 codons begins at a potential start triplet (ATG) 5 bases downstream from the *motB* stop codon. To confirm that this was the proper *cheA* reading frame, and to delineate the regions in which the start sites should lie, we determined the sequence changes in two *cheA* nonsense mutations. According to the two-start model (Fig. 1), amber mutation *cheA169* must lie between start(L) and

start(S) because it produces CheA<sub>S</sub> molecules of normal size, whereas *cheA140* must lie downstream of both start sites because it produces amber fragments of both CheA proteins (8). Both mutations create TAG triplets in the 654-codon open reading frame: *am169* at codon 10 and *am140* at codon 107 (Fig. 2). Thus, start(L) should be located between codons 1 and 10 of this open reading frame, and start(S) should be between codons 10 and 107.

**Location of *cheA* start sites.** Visual inspection of the pertinent portions of the *cheA* coding region revealed two potential translation starts at ATG triplets located at codons 3 and 98. Both are preceded by purine-rich sequences that could represent Shine-Dalgarno sites for initiating ribosome binding. However, when scanned with the W71 perceptron matrix used by Stormo et al. (10), which scores a variety of sequence features characteristic of orthodox translational starts, only the site at codon 98 had a positive score (+31).

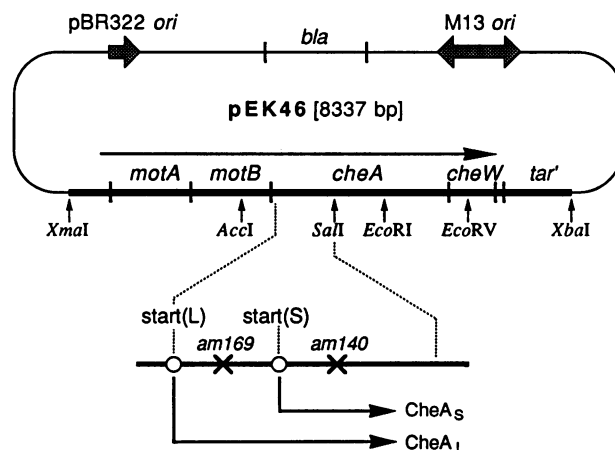


FIG. 1. Physical and genetic organization of the *cheA* region. Plasmid pEK46 is typical of those used in this project. Genes *motA* through *cheW* make up an operon of motility- and chemotaxis-related genes. Also shown are the relative positions of the two *cheA* translational start sites and two amber mutations (*cheA140* and *cheA169*) used to determine their correct reading frames. Other features indicated are as follows: *ori*, replication origin; *bla*,  $\beta$ -lactamase gene conferring resistance to ampicillin.

\* Corresponding author.

CAG	GTC	AGT	GTT	CCC	ACA	ATG	CCA	TCA	GCC	GAA	<u>CCG</u>	<u>AGG</u>	<u>TGA</u>	CAGC	(1)	<u>GTTG</u>	AGC	ATG	GAT	ATA	AGC	GAT	TTT	TAT	CAG	ACA	TTT	TTT	GAT	42		
														↓ <i>cheA</i> start (L)																		
														(1)	M	S	M	D	I	S	D	F	Y	Q	T	F	F	D	14			
GAA	GCG	GAC	GAA	CTG	TTG	GCT	GAC	ATG	GAG	CAG	CAT	TTG	CTG	GTT	TTG	CAG	CCG	GAA	GCG	CCA	GAT	GCC	GAA	CAA	TTG	AAT	GCC	ATC	TTT	132		
E	A	D	E	L	L	A	D	M	E	Q	H	L	L	V	L	Q	P	E	A	P	D	A	E	Q	L	N	A	I	F	44		
CGG	GCT	GCC	CAC	TCG	ATC	AAA	GGA	GGG	GCA	GGA	ACT	TTT	GGC	TTC	AGC	GTT	TTG	CAG	GAA	ACC	ACG	CAT	CTG	ATG	GAA	AAC	CTG	CTC	GAT	222		
R	A	A	H	S	I	K	G	G	A	G	T	F	G	F	S	V	L	Q	E	T	T	H	L	M	E	N	L	L	D	74		
GAA	GCC	AGA	CGA	GGT	GAG	ATG	CAA	CTC	AAC	ACC	GAC	ATT	ATC	AAT	CTG	TTT	TTG	GAA	ACG	<u>AAG</u>	<u>GAC</u>	ATC	↓ <i>cheA</i> start (S)		<u>ATG</u>	CAA	GAA	CAG	CTC	GAC	GCT	312
E	A	R	R	G	E	M	Q	L	N	T	D	I	I	N	L	F	L	E	T	K	D	I	M	Q	E	Q	L	D	A	104		
														T( <i>am140</i> )																		
TAT	AAA	CAG	TCG	CAA	GAG	CCG	GAT	GCC	GCC	AGC	TTC	GAT	TAT	ATC	TGC	CAG	GCC	TTG	CGT	CAA	CTG	GCA	TTA	GAA	GCG	AAA	GGC	GAA	ACG	402		
Y	K	Q	S	Q	E	P	D	A	A	S	F	D	Y	I	C	Q	A	L	R	Q	L	A	L	E	A	K	G	E	T	134		
CCA	TCC	GCA	GTG	ACC	CGA	TTA	AGT	GTG	GTT	GCC	AAA	AGT	GAA	CCG	CAA	GAT	GAG	CAG	AGT	CGC	AGT	CAG	TCG	CCG	CGA	CGA	ATT	ATC	CTT	492		
P	S	A	V	T	R	L	S	V	V	A	K	S	E	P	Q	D	E	Q	S	R	S	Q	S	P	R	R	I	I	L	164		
TCG	CCG	CTG	AAG	GCC	GGG	GAA	GTC	GAC	CTG	CTG	GAA	GAA	GAA	CTG	GGA	CAT	CTG	ACA	ACG	TTA	ACT	GAC	GTG	GTG	AAA	GGG	GCG	GAT	TCG	582		
S	P	L	K	A	G	E	V	D	L	L	E	E	E	L	G	H	L	T	T	L	T	D	V	V	K	G	A	D	S	194		
CTC	TCG	GCA	ATA	TTA	CCG	GGC	GAC	ATC	GCC	GAA	GAT	GAC	ATC	ACA	GCG	GTA	CTC	TGT	TTT	GTG	ATT	GAA	GCC	GAT	CAG	ATT	ACC	TTT	GAA	672		
L	S	A	I	L	P	G	D	I	A	E	D	D	I	T	A	V	L	C	F	V	I	E	A	D	Q	I	T	F	E	224		
ACA	GTA	GAA	GTC	TCG	CCA	AAA	ATA	TCC	ACC	CCA	CCA	GTG	CTT	AAA	CTG	GCA	GCC	GAA	CAA	GCG	CCA	ACC	GGC	GCG	GTG	GAG	CGG	GAA	AAA	762		
T	V	E	V	S	P	K	I	S	T	P	P	V	L	K	L	A	A	E	Q	A	P	T	G	R	V	E	R	E	K	254		
ACG	ACG	GCG	AGC	AAT	GAA	TCC	ACC	AGC	ATC	CGT	GTA	GCG	GTA	GAA	AAG	GTT	GAT	CAA	TTA	ATT	AAC	CTC	GTC	GGC	GAG	CTG	GTT	ATC	ACC	852		
T	T	R	S	N	E	S	T	S	I	R	V	A	V	E	K	V	D	Q	L	I	N	L	V	G	E	L	V	I	T	284		
CAG	TCC	ATG	CTT	GCC	CAG	CGT	TCC	AGC	GAA	CTG	GAC	CCG	GTT	AAT	CAT	GGT	GAT	TTG	ATA	ACC	AGC	ATG	GGG	CAG	TTA	CAA	CGT	AAC	GCC	942		
Q	S	M	L	A	Q	R	S	S	E	L	D	P	V	N	H	G	D	L	I	T	S	M	G	Q	L	Q	R	N	A	314		
CGT	GAT	TTG	CAG	GAA	TCA	GTG	ATG	TCG	ATT	CGC	ATG	ATG	CCG	ATG	GAA	TAT	GTT	TTT	AGT	CGC	TAT	CCC	CCG	CTG	GTG	CGT	GAT	CTG	GCG	1032		
R	D	L	Q	E	S	V	M	S	I	R	M	M	P	M	E	Y	V	F	S	R	Y	P	R	L	V	R	D	L	A	344		
GGA	AAA	CTC	GGC	AAG	CAG	GTA	GAA	CTG	ACG	CTG	GTG	GGC	AGT	TCT	ACT	GAA	CTC	GAC	AAA	AGC	CTG	ATA	GAA	GCG	ATT	ATC	GAC	CCG	CTG	1122		
G	K	L	G	K	Q	V	E	L	T	L	V	G	S	S	T	E	L	D	K	S	L	I	E	R	I	I	D	P	L	374		
ACC	CAC	CTG	GTA	GCG	AAT	AGC	CTC	GAT	CAC	GGT	ATT	GAA	CTG	CCA	GAA	AAA	CGG	CTC	GCC	GCA	GGT	AAA	AAC	AGC	GTC	GGA	AAT	TTA	ATT	1212		
T	H	L	V	R	N	S	L	D	H	G	I	E	L	P	E	K	R	L	A	A	G	K	N	S	V	G	N	L	I	404		
CTG	TCT	GCC	GAA	CAT	CAG	GGC	GGC	AAC	ATT	TGC	ATT	GAA	GTG	ACC	GAC	GAT	GGG	GCG	GGG	CTA	AAC	CGT	GAG	CGA	ATT	CTG	GCA	AAA	GCG	1302		
L	S	A	E	H	Q	G	G	N	I	C	I	E	V	T	D	D	G	A	G	L	N	R	E	R	I	L	A	K	A	434		
GCC	TCG	CAA	GGT	TTG	ACT	GTC	AGC	GAA	AAC	ATG	AGC	GAC	GAC	GAA	GTC	GCG	ATG	CTG	ATA	TTT	GCA	CCT	GGC	TTC	TCC	ACG	GCA	GAG	CAG	1392		
A	S	Q	G	L	T	V	S	E	N	M	S	D	D	E	G	R	C	L	I	F	A	P	G	F	S	T	A	E	Q	464		
GTC	ACC	GAC	GTC	TCC	GGG	GCG	GGC	GTC	GGC	ATG	GAC	GTC	GTT	AAA	CGT	AAT	ATC	CAG	AAG	ATG	GGC	GGT	CAT	GTC	GAA	ATC	CAG	TCG	AAG	1482		
V	T	D	V	S	G	R	G	V	G	M	D	V	V	K	R	N	I	Q	K	M	G	H	V	E	I	Q	S	K	494			
CAG	GGT	ACT	GGC	ACT	AGC	ATC	CGC	ATT	TTA	CTG	CCG	CTG	ACG	GTC	GCC	ATC	CTC	GAC	GGC	ATG	TCC	GTA	CGC	GTT	GCG	GAT	GAA	GTT	TTC	1572		
K	G	T	G	T	T	I	R	I	L	L	P	L	T	L	A	I	L	D	G	M	S	V	R	V	G	D	E	V	F	524		
ATT	CTG	CCG	CTG	AAT	GCT	GTT	ATG	GAA	TCA	CTG	CAA	CCC	CGT	GAA	GCC	GAT	CTC	CAT	CCA	CTG	GCC	GGC	GGC	GAG	CGG	GTG	CTG	GAA	GTG	1662		
I	L	P	L	N	A	V	M	E	S	L	Q	P	R	E	A	D	L	H	P	L	A	G	G	E	R	V	L	E	V	554		
CGG	GGT	GAA	TAT	CTG	CCC	ATC	GTC	GAA	CTG	TGG	AAA	GTG	TTC	AAC	GTC	GCG	GGC	GCG	AAA	ACC	GAA	GCC	ACC	CAG	GGA	ATT	GTG	GTG	ATC	1752		
R	G	E	Y	L	P	I	V	E	L	W	K	V	F	N	V	A	G	A	K	T	E	A	T	Q	G	I	V	V	I	584		
TTA	CAA	AGT	GGC	GGT	GCG	GCG	TAC	GCC	TTG	CTG	GTG	GAT	CAA	TTA	ATT	GGT	CAA	CAC	CAG	GTT	GTG	GTT	AAA	AAC	CTT	GAA	AGT	AAC	TAT	1842		
L	Q	S	G	G	R	R	Y	A	L	L	V	D	Q	L	I	G	Q	H	Q	V	V	V	K	N	L	E	S	N	Y	614		
CGC	AAA	GTC	CCC	GGC	ATT	TCT	GCT	GCG	ACC	ATT	CTT	GGC	GAC	GGC	AGC	GTG	GCA	CTG	ATT	GTT	GAT	GTC	TCC	GCC	TTG	CAG	GCG	ATA	AAC	1932		
R	K	V	P	G	I	S	A	A	T	I	L	G	D	G	S	V	A	L	I	V	D	V	S	A	L	Q	A	I	N	644		
														↓ <i>cheA</i> stop																		
CGC	GAA	CAA	CGT	ATG	GCG	AAC	ACC	GCC	GCC	TGA	ATGAGTAAAGGTTAACAAAT	↓ <i>cheW</i> start		ATG	ACC	GGT	ATG	ACG	AAT	GTA	ACA	AAG	CTG	GCC	AGC	GAG	2024					
R	Q	Q	R	M	A	N	T	A	A	•	(654)																					

FIG. 2. Sequence of the *cheA* gene. Important features of the nucleotide sequence are shown above the DNA sequence. Shine-Dalgarno sites and initiation codons are underlined; stop codons are indicated by a black bullet. The predicted primary structure of the *cheA* proteins is listed below the sequence, using the single-letter amino acid code. The nucleotide sequence is numbered from the first base of the first *cheA* codon. The amino acid residues in CheA<sub>L</sub> are also numbered.

The next highest value, -8, occurred at the GTG 6 bases upstream from the ATG at codon 3. When this GTG was replaced by ATG, the perceptron score increased to +83, implying that it might be a valid translational start. Few GTG initiation signals were used to train the W71 matrix, which

may account for the improvement following this substitution.

To establish the precise locations of the *cheA* translational initiation sites, we determined the N-terminal amino acid sequences of CheA<sub>L</sub> and CheA<sub>S</sub>. These proteins were pre-

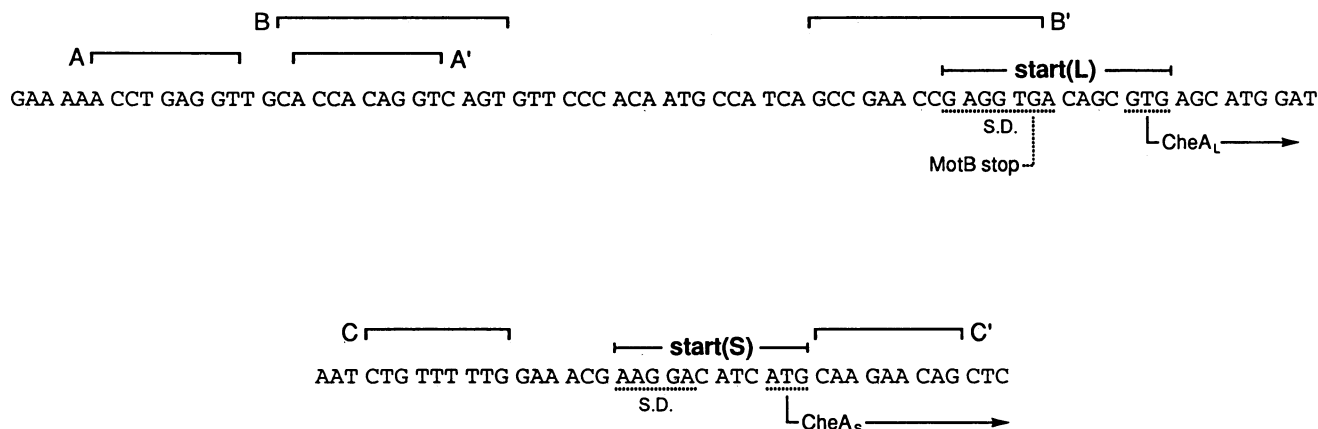


FIG. 3. Potential mRNA secondary structures and other sequence features in the vicinity of the two *cheA* start sites. Segments A/A', B/B', and C/C' make up inverted near repeats that could conceivably base pair to form hairpin structures in the mRNA. B/B' pairing would place the Shine-Dalgarno site of start(L) in a stem, whereas C/C' pairing would place the Shine-Dalgarno site and initiation codon of start(S) in a loop.

pared from plasmids pDV4 and pDV41, respectively, both of which contain the *cheA* coding region under the transcriptional control of an inducible *trp* promoter (3). A tract containing the translation initiation region and the first few codons of the *cheY* gene, which has an unusually efficient start site (4), precedes the insert. In pDV4, the *motB* fragment upstream of *cheA* is frame-shifted with respect to the *cheY* start and terminates early, whereas in pDV41 it is properly positioned and translated in frame. Upon induction with 100  $\mu$ g of 3- $\beta$ -indoleacrylic acid per ml, CheA<sub>L</sub> and CheA<sub>S</sub> are made from pDV4 in a ratio of roughly 20:1, whereas from pDV41 they are made in a ratio of about 5:1. Expression levels were determined by scanning densitometry of sodium dodecyl sulfate-polyacrylamide gel electrophoresis bands visualized with anti-CheA antiserum coupled to [<sup>35</sup>S]protein A. Although the absolute expression levels from the two plasmids are not strictly comparable owing to possible differences in copy number, promoter strength, and polarity effects, the CheA<sub>L</sub>/CheA<sub>S</sub> ratio should be insensitive to these variables. Thus, the different expression ratios in the two plasmids must reflect differences in the relative rates of initiation at the two start sites, which could be subject to physiological regulation, as discussed next.

The proteins were purified electrophoretically and their amino-terminal sequences were determined by automated Edman degradation (2). CheA<sub>S</sub> began with M Q E Q L D, which agrees unambiguously with the predicted sequence following the putative start(S) initiation site at codon 98. The terminus of CheA<sub>L</sub> resembled two superimposed sequences, the major being S M D I S D and the minor being M D I S D F. This result can be explained by initiation of translation at codon 1, followed by proteolytic processing at the N terminus. This would yield a mixed population lacking the initial methionine (encoded by GTG) and partially lacking the subsequent serine. We conclude that translation of CheA<sub>L</sub> is initiated primarily at codon 1, but these data do not preclude use of the ATG at codon 3 as a minor, unorthodox initiation site. Stock et al. (9) observed the same N-terminal heterogeneity in CheA<sub>L</sub> of *Salmonella typhimurium*. Note that this organization places the termination codon of *motB* within the Shine-Dalgarno tract of start(L).

**Possible translational control of *cheA* expression.** The different expression patterns of plasmids pDV4 and pDV41 (see

above) imply that *cheA* may be subject to several kinds of translational control. First, CheA<sub>L</sub> expression is about fourfold lower from pDV41, in which ribosomes traverse the upstream *motB* sequence in frame, than from pDV4, in which translation of the sequence upstream of *cheA* is terminated by a shift in reading frame. This difference in CheA<sub>L</sub> expression implies that *motB* translation may interfere with initiation at start(L). Conceivably, the rather unusual placement of the *motB* stop codon within the Shine-Dalgarno tract of start(L) (Fig. 3) could lead to an inhibitory effect of this sort. This situation contrasts with the overlap of the *motA* stop codon and the *motB* initiation codon, which is thought to cause translational coupling of these two genes (1).

A second control mechanism may regulate the relative expression of CheA<sub>L</sub> and CheA<sub>S</sub>, whose levels appear to be reciprocally related. In pDV4, CheA<sub>L</sub> expression is relatively high and the amount of CheA<sub>S</sub> is low. In pDV41, in which CheA<sub>L</sub> levels are reduced, CheA<sub>S</sub> expression is elevated. Since the mRNA transcribed from the *cheA* operon does not appear to undergo endonucleolytic processing (data not shown), both CheA products are probably translated from identical mRNA molecules. The sequences surrounding the two *cheA* start sites reveal potential mRNA secondary structures that could play a role in modulating their initiation rates (Fig. 3). Two mutually exclusive hairpins (A/A', B/B') could form near start(L), one of which (B/B') would obscure the *motB* stop and start(L) ribosome binding site. Another potential hairpin (C/C') embraces start(S), but, unlike the one at start(L), it should serve to expose the Shine-Dalgarno region and initiation codon. If start(L) normally captures most of the ribosomes moving down the message, translation at start(S) would depend on attraction of free ribosomes. However, ribosomes emanating from start(L) would be expected to disrupt the C/C' hairpin and occlude start(S), hindering expression of CheA<sub>S</sub>. Thus, the efficiency of initiation at start(L) should dictate the rate of initiation at start(S), and the expression rates of CheA<sub>L</sub> and CheA<sub>S</sub> should be inversely related. Whether these regulatory effects play a role in fine-tuning the chemotaxis machinery of *E. coli* to different physiological conditions remains to be determined.

**Nucleotide sequence accession number.** The *cheA* sequence

has been submitted to GenBank under accession number M34669.

This work was supported by Public Health Service research grant GM28706 from the National Institutes of Health.

We thank Bob Bourret for carefully scrutinizing early versions of our *cheA* sequence.

#### REFERENCES

1. Dean, G. E., R. M. Macnab, J. Stader, P. Matsumura, and C. Burks. 1984. Gene sequence and predicted amino acid sequence of the MotA protein, a membrane-associated protein required for flagellar rotation in *Escherichia coli*. *J. Bacteriol.* **159**:991–999.
2. Edman, P., and G. Begg. 1967. A protein sequenator. *Eur. J. Biochem.* **1**:80–91.
3. Matsumura, P. Unpublished data.
4. Matsumura, P., J. J. Rydel, R. Linzmeier, and D. Vacante. 1984. Overexpression and sequence of the *Escherichia coli cheY* gene and biochemical activities of the CheY protein. *J. Bacteriol.* **160**:36–41.
5. Parkinson, J. S., and S. E. Houts. 1982. Isolation and behavior of *Escherichia coli* deletion mutants lacking chemotaxis functions. *J. Bacteriol.* **151**:106–113.
6. Sanger, F., S. Nicklen, and A. R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **74**:5463–5467.
7. Smith, R. A. 1981. Detailed analysis of a genetic locus that contains a pair of overlapping genes and is involved in bacterial chemotaxis. Ph.D. thesis. University of Utah, Salt Lake City.
8. Smith, R. A., and J. S. Parkinson. 1980. Overlapping genes at the *cheA* locus of *E. coli*. *Proc. Natl. Acad. Sci. USA* **77**:5370–5374.
9. Stock, A., T. Chen, D. Welsh, and J. Stock. 1988. CheA protein, a central regulator of bacterial chemotaxis, belongs to a family of proteins that control gene expression in response to changing environmental conditions. *Proc. Natl. Acad. Sci. USA* **85**:1403–140.
10. Stormo, G. D., T. D. Schneider, L. Gold, and A. Ehrenfeucht. 1982. Use of the "perceptron" algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Res.* **10**:2997–3011.
11. Vieira, J., and J. Messing. 1987. Production of single-stranded plasmid DNA. *Methods Enzymol.* **153**:3–34.