Volume 10 Number 22 1982 Nucleic Acids Research

Sequence of the *malK* gene in *E.coli* K12

Eric Gilson, Hiroshi Nikaido* and Maurice Hofnung

Unité de Programmation Moléculaire et Toxicologie Génétique (CNRS LA 271 - INSERM U.163) Institut Pasteur, 28 rue du Dr.Roux, 75015 Paris, France and *Department of Microbiology and Immunology, University of California, Berkeley, CA 94720, USA

Received 5 August 1982; Revised and Accepted 25 October 1982

ABSTRACT

We present the sequence of gene malK which encodes a component of the system for maltose transport in E.coli K12. We also determined the position of deletion (S50) which fuses malK to the following gene lamB ; the malK-lamB protein hybrid contains all of the malK protein. The mRNA corresponding to the last two thirds of gene malK could form stable stem and loop structures. The malK protein, as deduced from the gene sequence, would include 370 residues and correspond to a molecular weight of 40700. The sequence as well as sequence comparisons with the ndh protein of E.coli are discussed in terms of the location and function of the malK protein.

INTRODUCTION

All of the known binding protein dependent transport systems in bacteria appear to require several proteins (1). The molecular mechanisms underlying such systems are not yet well understood. The active transport system for maltose in E.coli K12 includes at least five proteins encoded in the genes of the malB region i.e. the malK-lamB and malE-malF-malG operons (2). All these proteins, except the malG product, have been identified. It is generally believed that the lamB protein forms a partially specific transmembrane channel which facilitates maltose and maltodextrins diffusion through the outer membrane (3,4). The early steps of transport also involve an interaction between the lambda receptor and the maltose binding protein (3). Subsequently, the malE gene product (5) would be involved in the concentration of the substrate into the periplasmic space (6) and would allow their capture by a group of proteins located at the level of the inner membrane. This group is likely to comprise the products of the malF, malK and malG genes (7,8,9). One hypothesis proposes that these inner membrane components of the system constitute an energy dependent pore which would allow the translocation of the substrate through the cytoplasmic membrane (10,11). However, nothing is known on the exact enzymatic function of these inner membrane components including the malK protein.

We present here the complete nucleotide sequence of the malK gene. We discuss briefly the results in terms of the structure of the malK-lamB operon and of the possible location and function of the malK protein.

0305-1048/82/1022-7449$ 2.00/0

MATERIALS AND METHODS

E.coli strain pop 2044 (12) was used for plasmid propagation. The plasmids used for the isolation of DNA fragments were $pHCP_2$ and $pHCP_3$ (13) carrying respectively the EcoR I (0)-Bgl II (+3168) and EcoR I (0)-Sal I (+993) fragments of the malB region cloned in pBR322. Deletion S50 is carried by the phage lambda aph80malB13S50 (14). Plasmid DNAs were purified from clear lysates by using one cesium chloride gradient centrifugation (15) followed by one 5% to 20% sucrose gradient in Tris HCl 10^{-2}M pH = 7,4, E.D.T.A. 10^{-3}M, NaCl 10^{-1}M. The phage DNA was purified as described (16).

All enzymes used were bought from Biolabs and Boehringer. Reactions were carried out under conditions described by suppliers.

DNA fragments were separated by polyacrylamide gel electrophoresis and isolated by electroelution followed by a step on a DEAE cellulose column. The restriction map was established by partial hydrolysis of labelled fragments (17).

DNA fragments were 5' labelled with (gamma^{32}P) by the kinase exchange reaction (18) and sequenced as described by Maxam and Gilbert (19).

The EcoR I (0)-Hinf I (+1078) fragment which we have sequenced was obtained from a functional malB region (13). In addition several lines of evidence confirm that it corresponds to a functional malK gene. Firstly, the nucleotide sequence data presented here correspond well to the size and the position expected for the malK gene. Secondly, deletion S50 verifies the translational phase at the end of the malK. Thirdly, the molar ratio Tyr to Arg deduced from the malK gene sequence is in good agreement with that of the malK protein (9).

RESULTS AND DISCUSSION

In order to determine the complete nucleotide sequence of the malK gene, we have sequenced a DNA fragment between the EcoR I (0) site and the Hinf I (+1078) site of the malB region (Figure 1). The sequences at the left of EcoR I (0)(20) and at the right of Sal I (+993) (21) were already known. Plasmid $pHCP_3$ (Materials and Methods) was used for the determination of the fine restriction map while both plasmids $pHCP_2$ and $pHCP_3$ were used for the determination of the sequence. Figure 1 shows the map and the strategy we adopted for sequencing this region. The restriction map was confirmed by DNA sequencing except for one Taq I (+657) site (TCGA) which had not been found by the mapping. This site corresponds to the recognition sequence of the dam$^+$ function (G^{m}ATC) (22) and thus was presumably protected against the Taq I enzyme.

Since the EcoR I (0) and the Sal I (+993) sites are expected to be located within the malK gene (20,14), we have looked for one reading frame beginning on the left side of the EcoR I (0) site and ending on the right side of the Sal I (+993) site. Indeed, there is an open phase from nucleotide -194 to +1021 (Figure 2). Three ATGs and one GTG

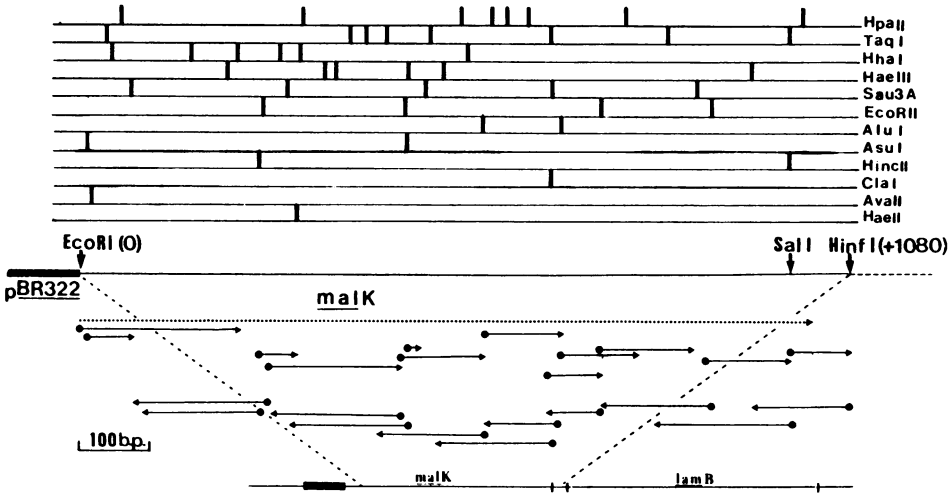

Figure 1. Restriction map and sequence strategy for gene malK.
The locations of restriction sites (Materials and Methods) are displayed at the top of the figure. The middle part of the figure shows the sequencing strategy with the region of the plasmid pHCP2 used in this work. The nucleotides are numbered from the EcoR I (O) site in malK. Arrows indicate the direction and extent of sequence obtained from the labelled end of each fragment (●). In the bottom part, the malK-lamB operon of the malB region is represented. The heavy box 5' to malK represent the regulatory region of the malB operon.

belonging to this phase are possible translation starts for malK (20). However, it has been determined that the 5' end of the malK-lamB mRNA is located after the GTG (-182) and ATG (-158) but before the ATGs (-137) and (-92) (23). Of the latter, the ATG (-92) which is preceded by a sequence complementary to 3' end of the 16S rRNA sequence (24) is the most likely translational start for malK (Figure 2).

To confirm the phase at the end of malK we have determined the localization of deletion S50 which results in the formation of a malK-lamB protein fusion (25). In this fusion malK and lamB are expected to be in the same translational phase. The sequencing results (Figure 3) fulfilled this expectation, thus confirming that the open reading frame corresponds to the malK gene. It is worth mentioning that deletion S50 fuses the malK gene to the lamB gene within the stop codon TAA of the malK gene (Figure 3). Thus, the hybrid protein malK-lamB should contain all the amino-acids of the malK protein.

Special features of the sequence

Genetic evidence suggested the existence of a low level promoter located before the distal structural gene lamB (12,13) in the malK-lamB operon. A search for a DNA segment containing homology to the -35 and -10 regions known as consensus promoter

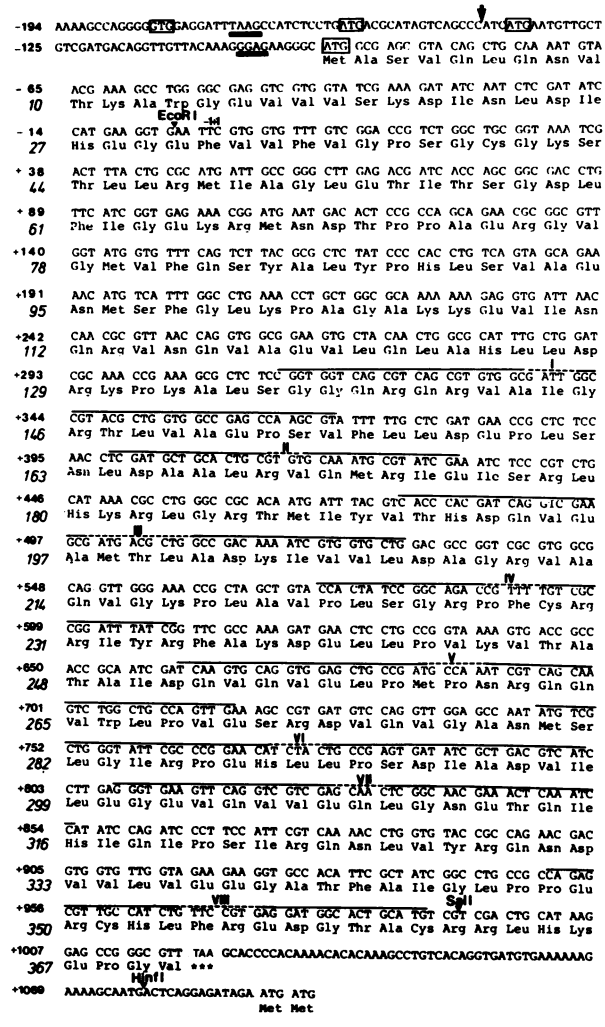

Figure 2. The nucleotide sequence of gene malK and its flanking regions
The sequence extends from the first nucleotide of the malK open reading frame (- 194) to the second codon of the lamB gene. The sequence at the left of the EcoR I (0) site was determined previously (20) including the sequence through the EcoR I (0) site (H. Bedouelle, unpublished data). The three ATG's and the one GTG, which are possible initiation codons for malK, are boxed. The heavy lines at positions -170 and -99 indicates sequences complementary to the 3'OH end of the 16S rRNA. Transcription initiation occurs at position -140 (23). The amino-acid translation is shown from the ATG (-92) to position (+1021).
The thin lines above the nucleotide indicate palindromic sequences in the malK gene. The calculated "delta Gs" of these structures are : I -25 Kcal/mole ; II -14,6 Kcal/mole ; III -22,5 Kcal/mole ; IV -18,1 Kcal/mole ; V -14,4 Kcal/mole ; VI -9,3 Kcal/mole ; VII -4,3 Kcal/mole ; VIII -8,9 Kcal/mole. The dashed lines represent the loop of each structure.

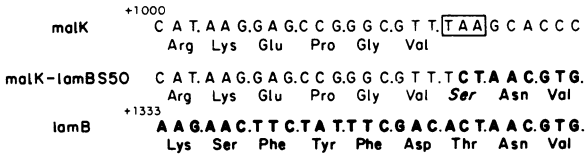

Figure 3. Position of deletion S50
The upper and lower lines display the wild type DNA sequence at the end of the malK gene and a part of the lamB gene starting at nucleotide (+1333). The middle line is the sequence corresponding to deletion S50. The deletion removes the nucleotides from +1019 to +1351.

sequences in E.coli (26) was carried out. However, no such sequence was found in the DNA from the EcoR I (0) site to the beginning of the lamB gene.

We have also looked for chi sites, known to increase recombinational activity in lambda and E.coli (27). A region homologous to the GCTGGTGG octamer, necessary for the Chi$^+$ phenotype was found in the DNA sequence of malK gene from position (+350) to (+358).

The codon usage in the malK gene is shown on Table 1. The pattern is not random : most of the codons corresponding to major tRNA species are preferently used (Leu, Pro, Lys, Arg, Ile, Gly), while for a few codons this preference is not observed (Ala, Val, Ser, Gly). Such a codon choice may be characteristic of intermittently or moderately expressed genes in E.coli (28).

The malK messenger RNA

An interesting feature of the product of transcription of the malK gene is the presence of a potentially stable stem and loop structure in the middle of the gene followed by a stretch of five uridine residues (Figure 4). This structure is analogous to a

Table 1. codon usage in the malK gene

UUU-PHE 5	UCU-SER 2	UAU-TYR 2	UGU-CYS 2
UUC-PHE 5	UCC-SER 5	UAC-TYR 3	UGC-CYS 2
UUA-LEU 1	UCA-SER 2	UAA- . 1	UGA- . 0
UUG-LEU 3	UCG-SER 3	UAG- . 0	UGG-TRP 2
CUU-LEU 2	CCU-PRO 2	CAU-HIS 7	CGU-ARG 12
CUC-LEU 8	CCC-PRO 1	CAC-HIS 2	CGC-ARG 10
CUC-LEU 4	CCA-PRO 6	CAA-GLN 9	CGA-ARG 1
CUG-LEU 22	CCG-PRO 12	CAG-GLN 13	CGG-ARG 3
AUU-ILE 7	ACU-THR 4	AAU-ASN 5	AGU-SER 1
AUC-ILE 14	ACC-THR 4	AAC-ASN 7	AGC-SER 4
AUA-ILE 0	ACA-THR 2	AAA-LYS 14	AGA-ARG 1
AUG-MET 10	ACG-THR 4	AAG-LYS 1	AGG-ARG 0
GUU-VAL 7	GCU-ALA 5	GAU-ASP 11	GGU-GLY 10
GUC-VAL 9	GCC-ALA 9	GAC-ASP 6	GGC-GLY 13
GUA-VAL 8	GCA-ALA 6	GAA-GLU 15	GGA-GLY 2
GUG-VAL 19	GCG-ALA 8	GAG-GLU 11	GGG-GLY 2

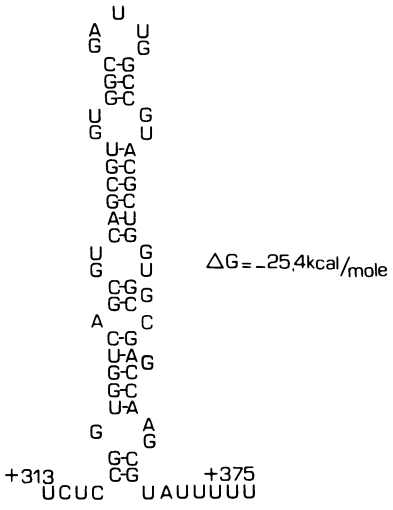

Figure 4. A terminator like potential stable RNA secondary structure in malK The "delta G" of this stem and loop structure (I in Figure 2) is -25 Kcal/mole. Numbers indicate the positions as presented in the DNA sequence (Figure 2).

classical Rho-independant termination site for transcription (29). It is followed but not preceded by a series of stable potential secondary structures (Figure 2). Such secondary structures have been implicated in the stabilization of mRNA and/or the control of translation for membrane proteins (30). Another although less likely possibility is that they play a rôle in the differential expression of the malK and lamB genes. Indeed the lambda receptor is produced in larger amount, up to 10^5 copies per cell (31), than the malK gene product, between 10^3 (8) and 10^4 (9) copies per cell. For example, these secondary structures of the RNA could reduce the efficiency of translation and/or transcription in certain conditions. Under these conditions, the secondary promoter (12,13) could be unmasked and thus allows an efficient transcription of the lamB gene.

The malK protein

The coding sequence beginning at the position -92 and ending at the position +1021 corresponds to a polypeptide of 370 amino-acids. The sequence of this protein is shown on Figure 2 with the nucleotide sequence. The molecular weight calculated from the sequence is 40700 daltons. This size corresponds well to the values of 40000 and 43000 daltons determined by polyacrylamide gel electrophoresis under denaturing conditions (8,9). Furthermore the molar ratio of Tyr to Arg in the malK protein deduced from the nucleotide sequence (0.19) and the one estimated in vivo (0.24) (9) are in good agreement.

From the predicted amino-acid sequence of the malK protein, a number of observations can be made concerning its function and localization.

A computer search has suggested a potential internal homology of a NH_2 terminal

```
258 Pro Met Pro Asn --- Arg --- --- Gln Gln Val Trp Leu Pro ---
 71 Pro --- Pro Ala Glu Arg Gly Val Gly Met Val Phe Gln Ser Tyr

269 Val Glu --- Ser Arg Asp Val Gln Val Gly Ala Asn Met Ser Leu
 85 Ala Leu Tyr Pro His --- Leu Ser Val Ala Glu Asn Met Ser Phe

283 Glv Ile Arg Pro Glu His Leu Leu Pro Ser Asp Ile Ala Asp ---
 99 Gly Leu Lys Pro Ala Gly Ala Lys Lys Glu Val Ile Asn Gln Arg

297 Val Ile Leu Glu Gly Glu Val Gln Val Val Glu Gln Leu Gly Asn
114 Val Asn Gln Val Ala Glu Val Leu Gln Leu Ala His Leu Leu Asp

312 Glu Thr Gln Ile His Ile Gln Ile Pro Ser Ile Arg Gln Asn Leu
129 Arg Lys Pro Lys Ala Leu Ser --- Gly Gly Gln Arg Gln Arg Val

327 Val Tyr ---Arg Gln Asn Asp Val Val Leu Val Glu Glu Gly Ala
143 Ala Ile Gly Arg --- Thr --- --- --- Leu Val Ala Glu Pro Ser

341 Thr Phe --- --- --- --- Ala Ile Gly --- Leu Pro Pro Glu ---
154 Val Phe Leu Leu Asp Glu Pro Leu Ser Asn Leu Asp Ala Ala Leu

350 Arg Cys His Leu --- Phe Arg Glu Asp Gly Thr Ala Asp Arg Arg
169 Arg Val Gln Met Arg Ile --- Glu --- --- Ile Ser --- --- Arg

364 Leu His Lys Glu Pro Gly Val
179 Leu His Lys Arg Leu Gly Arg
```

Figure 5. Possible internal homology in the malK protein
The three short, homologous segments (residues 92-102 vs.276-286, residues 148-155 vs.335-342, and residues 178-184 vs.363-369) are approximatively in register with each other. When the alignment score (37,38) was calculated for the match involving residues 92-185 and residues 276-370, as well as 20 matches of the same sequences randomized by computer, the difference between the score of the real sequences and the average score of random matches was found to be 4.4 times larger than the standard deviation of the random matches. This "alignment index" of 4.4 means that the probability of this match occuring by chance is of the order of 10^{-5}.

region with the CO_2 H terminal region. An alignment was generated as shown on Figure 5.

Membrane fractionation experiments established that the malK protein is bound to the inner membrane (8,9). Genetic evidence suggested that it could be peripherally bound to the membrane by means of the malG protein (8). The fact that the malK gene does not encode a typical NH_2 terminal signal sequence (23) is not very informative with respect to the localization of the malK protein since most known inner membrane proteins are not made with an NH_2 terminal signal peptide (33). The malK protein includes 42 acidic amino-acids (Glu, Asp ; Table 1)and 43 basic amino-acids (Arg, Lys ; Table 1). The average length of peptides devoid of such amino-acids is thus (370/85 =4.35) amino-acids residues. The distribution of these polar amino-acids is not random. On one hand there are 20 occurences of two consecutive such polar amino-acids. On the other hand there are a few continuous stretches of apolar amino-acids with more than 10 residues. The largest includes 17 residues (residues 76 to 93) : its average hydrophobicity index calculated according to Segrest and Feldmann is 2.1 (34). Shorter stretches, residues 31 to 41, residues 274 and 284 and residues 339 to 348 have average indices of 1.2, 1.9 and 1.5 respectively. All these average hydrophobicity indices fall within the triangle determined for apolar peptide from soluble proteins, while indices of apolar membrane spanning peptides usually fall outside (35). Thus the

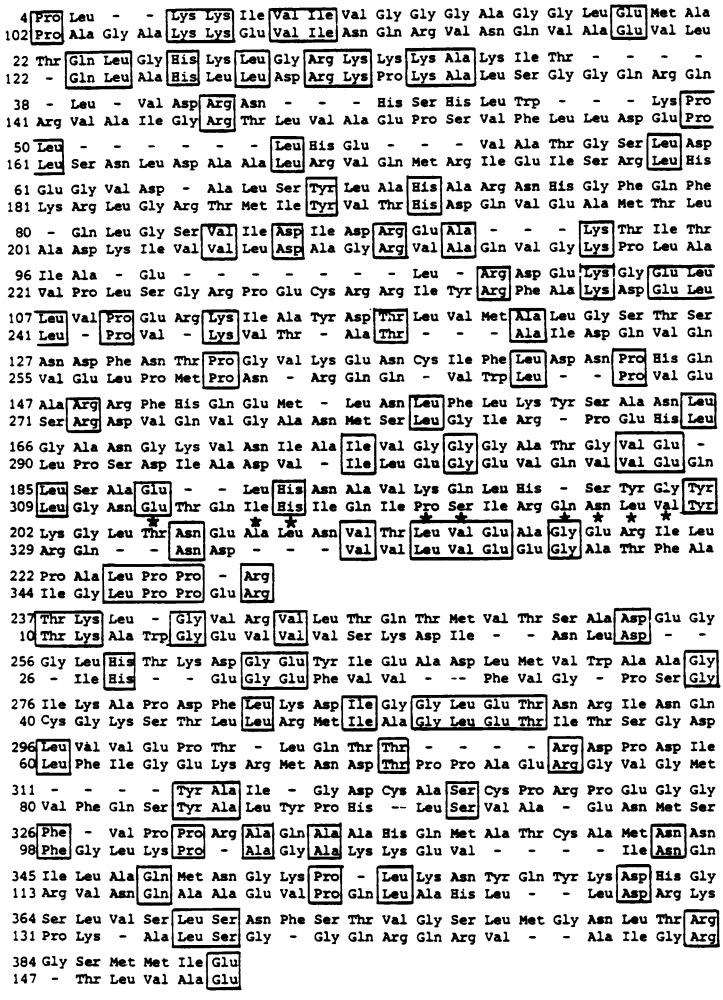

Figure 6. Sequence homologies between the malK protein and the ndh protein
The ndh protein (434 residues) is shown on lines 1, 3, 5... while the malK protein (370 residues) is shown on lines 2, 4, 6... The first alignment (upper part of the figure) corresponds to homology between residues 4-227 of the ndh protein and residues 102-350 of the malK protein. In this alignment 22 % of the positions are identical and 30 % are occupied by functionnally related amino-acids residues (37). The alignment index" (37,38) for the match involving residues 4-34 of the ndh protein and residues 102-133 of the malK protein is 3,93. The stars indicate the residues of the ndh protein which are homologous to the segment 181-195 of the hisP protein. This segment of the hisP protein was found to be homologous to segment 161-174 of the malK protein (37). The second alignement (lower part of the figure) corresponds to homology between the residues 237-389 of the ndh protein and residues 10-151 of the malK protein. In this alignment 20 % of the positions are identical and 30 % are occupied by functionally similar amino-acids residues (37). The main regions of matches were also detected using the Needleman and Wunsch algorithm (40) on the whole protein sequence although the final alignments presented differences in the regions of low homology.

predicted sequence of malK protein would be compatible with that of a peripheral membrane protein. However we can certainly not exclude on this basis that the malK protein could span the membrane. Indeed homology was found between the malK protein and the hisP protein which has been assumed to span the membrane (36). This homology will be presented elsewhere (37).

One hypothesis on the rôle of the malK protein proposes that it could function as an ATPase (10) allowing "energization" of the malF and malG proteins acting as a largely unspecific pore. We have compared the sequence of the malK protein with that of the E.coli F_1 - ATPase subunit alpha (32) but found no significant homology which would have substantiated this hypothesis. However we have found significant homology between the sequence of the malK protein and that of the respiratory NADH deshydrogenase of Escherichia coli (the ndh protein) (39), as shown on Figure 6. This raises the interesting possibility that the malK protein may play a rôle in energization through a mechanism involving an oxydo reduction reaction rather than ATP hydrolysis. This homology may also (or alternatively) correspond to other similarities between the malK protein and the ndh protein, such as membrane localization or interaction with a common component. Finally such sequence homologies may have no functional significance but reflect only the existence of a common ancestor to both proteins.

It may be relevant to mention on this respect that a comparison between the ndh and the hisP protein revealed only homology between the regions of residues 205-220 of the ndh protein and that of residues 181-195 of the hisP protein (Figure 6). However each of these regions is homologous to a different region in the malK protein. Thus, conclusions concerning the implications of sequence homologies should be taken with caution in absence of further experimental data.

ACKNOWLEDGMENTS

We thank Jean Marie CLEMENT for showing one of us (E.G.) DNA technology and for providing some DNA samples, Hugues BEDOUELLE for communication of unpublished data, Doug BRUTLAG for help with the computer and Intelligenetics for use of their experimental programs. This work was supported by grants from the Centre National de la Recherche Scientifique (C.P. 960002), North Atlantic Treaty Organization (grant 1297), the Fondation pour la Recherche Médicale to M.H., and from the N.I.H. (AI 09664) to H.N.

REFERENCE

1. Dills, S.S., Apperson, A., Schmidt, M.R., Saier, M.H.Jr. (1980) Microbiol. Rev. 44, 385-418.
2. Raibaud, O., Roa, M., Braun-Breton, C. and Schwartz, M. (1979) Mol. Gen. Genet. 174, 241-248.
3. Ferenci, T. and Boos, W. (1980) J. Supramol. Structure, 13, 101-116.

4. Nikaido, N. Luckey, M. and Rosenberg, E.Y. (1980) J. Supramol. Structure, 13, 305-313.
5. Kellerman, O., Szmelcman, S. (1974) Eur. J. Biochem. 47, 139-140.
6. Silhavy, T.J., Szmelcman, S., Boos, W. and Schwartz, M. (1975) Proc. Natl. Acad. Sci. USA, 72, 2120-2124.
7. Shuman, H.A., Silhavy, T.J., Beckwith, J. (1980) J. Biol. Chem. 25, 168-174.
8. Shuman, H.A., Silhavy, T.J. (1981) J. Biol. Chem. 256, 560-562.
9. Bavoil, P., Hofnung, M., Nikaido, H. (1980) J. Biol. Chem. 255, 8366-8369.
10. Boos, W. (1982) Annales de Microbiol. (Inst. Pasteur) 133A, 145-162.
11. Shuman, H.A. (1982) Annales de Microbiol. (Inst. Pasteur) 133A, 153-162.
12. Braun-Breton, C. and Hofnung, M. (1978) Mol. Gen. Genet. 159, 143-149.
13. Clément, J.M., Perrin, D. and Hedgpeth, J. (1982) Mol. Gen. Genet. 185, 302-310.
14. Raibaud, O., Clément, J.M., Hofnung, M. (1979) Molec. Gen. Genet. 174, 261-267.
15. Davis, R.W., Botstein, D., Roth, J.R. (1980) Cold Spring Harbor Laboratory. 116-119.
16. Marchal, C., Greenblatt, J. and Hofnung, M. (1978) J. Bacteriol. 136, 1109-1119.
17. Smith, H.O. and Birnstiel, M. (1976) Nucl. Acids, Res. 3, 2387-2398.
18. Berkner, K. and Folk, W. (1977) J. Biol. Chem. 252, 3176-3184.
19. Maxam, A. and Gilbert, W. (1980) Meth. Enzymol. 65, 499-560.
20. Bedouelle, H. and Hofnung, M. (1982) Mol. Gen. Genet. 185, 82-87.
21. Clément, J.M. and Hofnung, M. (1981) Cell, 27, 507-514.
22. Mc Clelland, M. (1981) Nucl. Acids, Res. 9, 5859-5866.
23. Bedouelle, H. Schmeissner, U., Hofnung, M., Rosenberg, M. (1982) J. Mol. Biol. in press.
24. Shine, J. and Dalgarno, L. (1974) Proc. Natl. Acad. Sci. USA, 71, 1342-1346.
25. Emr, S.D. and Silhavy, T.J. (1980) J. Mol. Biol. 141, 63-90.
26. Rosenberg, M., Court, D. (1979) Annu. Rev. Genet. 13, 319-353.
27. Horwitz, H., Christie, G.E., Platt, T. (1982) J. Mol. Biol. 156, 245-256.
28. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. (1981) Nucl. Acids, Res. 9, r43-r74.
29. Platt, T. (1981) Cell, 24, 10-23.
30. Movva, N.R., Nakamura, K. and Inouye, M.(1980) J. Mol. Biol. 143, 317-328.
31. Braun, V. and Krieger-Brauer, H.J. (1977) Biochem. Biophys. Acta, 469, 89-98.
32. Gay, N.J. and Walker, J.E. (1981) Nucl. Acids, Res. 9, 2187-2194.
33. Higgins, C.F., Haag, P.O., Nikaido, H., Ardeshir, F., Garcia, G. and Ferro-Luzzi Ames, G. (1982) Nature, 298, 723-727.
34. Segrest, J.P. and Feldmann, R.J. (1974) J. Mol. Biol. 87, 853-858.
35. Clément, J.M. (1982) Thèse de Doctorat d'Etat. Paris VI.
36. Ferro-Luzzi Ames, G. and Nikaido, H. (1978) Proc. Natl. Acad. Sci. USA, 75, 5447-5451.
37. Gilson, E., Higgins, C.F., Hofnung, M., Ferro-Luzzi Ames, G. and Nikaido, H. (1982) J. Biol. Chem. 257, 9915-9918.
38. Barker, W.C. and Dayhoff, M.O. (1972) in Atlas of protein sequence and structure 1972 (Dayhoff, M.O. ed), 5, 101-110, National Biomedical Research Foundation, Washington D.C.
39. Young, I.G., Rogers, B.L., Campbell, H.D., Jaworowski, A. and Shaw, D.C. (1981) Eur. J. Biochem. 116, 165-170.
40. Needleman, S.D. and Wunsch, C.D. (1970) J. Mol. Biol. 48, 443-453.