Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 6602–6606, September 1988
Biochemistry

A large family of bacterial activator proteins

(protein homology/DNA binding motif/LysR family)

STEVEN HENIKOFF*†, GEORGE W. HAUGHN‡§, JOSEPH M. CALVO‡, AND JAMES C. WALLACE*

*Fred Hutchinson Cancer Research Center, 1124 Columbia Street, Seattle, WA 98104; and ‡Section of Biochemistry, Molecular and Cell Biology, Cornell University, Ithaca, NY 14853

Communicated by H. E. Umbarger, May 2, 1988

ABSTRACT **At least nine different bacterial proteins belong to the LysR family. The gene sequence for one of these proteins is presented here. Six others (*Escherichia coli* LysR, IlvY, CysB; *Salmonella typhimurium* MetR; *Rhizobium* NodD; and *Enterobacter cloacae* AmpR) are known to activate other genes. Based on sequence alignments, each member of this family is predicted to have a helix-turn-helix DNA binding motif near its amino terminus. The combined evidence indicates that all nine proteins are related by common ancestry, are similarly folded, and are not detectably related to other known bacterial regulatory proteins. The DNA database searching procedure and other methods used in this study should be useful in detecting other groups of related proteins.**

A full understanding of gene expression will require the elucidation of mechanisms whereby different regulatory proteins interact with specific DNA sequences. This task has been simplified in recent years by the discovery of short peptide motifs within diverse proteins that interact with DNA. One is the helix-turn-helix motif originally seen in λ Cro protein and later recognized in many other DNA-binding proteins (1, 2). A second is the zinc finger domain, originally hypothesized for *Xenopus* transcription factor IIIA and later recognized by sequence comparisons to be in several other zinc-dependent DNA-binding proteins (3). These short motifs are found in regulatory proteins that might not be similar overall. However, in some cases, DNA-binding proteins resemble one another more extensively, implying common ancestry and a similar mechanism of action (4–8).

Here we report extensive similarities between the sequences for nine bacterial proteins, six of which are known to activate other genes. These relationships were found initially by searching DNA databases for amino acid sequence homologies. The relationships were confirmed by completion of the sequence of one family member¶ and by pairwise and consensus comparisons. Each protein was scored for the likelihood that it contains a helix-turn-helix motif. Evidence for a helix-turn-helix motif in each case at an aligned position argues that all members of the family are helix-turn-helix DNA-binding proteins.

METHODS

Database searches, dot matrices, and alignments were done with GENEPRO version 4.1 software and EMBL 14, GenBank 52, and NBRF-PIR 15 databases obtained from Riverside Scientific Enterprises (Seattle). Searches were done on a BIOS AT personal computer obtained from Lang Systems (Arlington, MA).

The DNA database searching procedure involved comparing a "query sequence" to every possible translated reading frame of every database sequence (9). The software that we used performed these operations by fetching an individual nucleotide sequence, translating each reading frame into protein, comparing that reading frame with the query, repeating the comparison for the next reading frame, and then repeating the entire operation for the next nucleotide sequence.

The comparison strategy was to align 90 amino acids of the query with a stretch of 90 amino acids from a database sequence and calculate a logarithm of odds (lod) score, which measures the likelihood that two aligned amino acids are functionally equivalent (10). The query was then aligned with the next stretch of translated sequence and a lod score was calculated. To reduce the number of unproductive alignments, only sequences in which one or more dipeptides match between the probe and the translated database sequences were considered (11).

The dot matrix procedure (12) compared all segments of 100 amino acids, placing a dot whenever 15 or more matches were found and a lod score of at least 75 was obtained. A window of 100 residues was also used for computing pairwise alignments, using the lod matrix with a gap penalty of 10, a size penalty of 3, and a maximum allowable gap of 5.

Multiple alignments were performed by a procedure similar to that recommended by others (13, 14). Starting with the best pairwise alignment, a third sequence was added, introducing gaps as appropriate. This was repeated for the remaining sequences. Refinement involved manual adjustment to remove and cluster gaps.

A consensus sequence was arrived at by choosing the most frequent residue at each position, with the following exceptions: (*i*) A residue present only twice at a position was chosen when at least one of the other aligned residues was chemically similar (similar amino acids are Ser, Thr, Pro, Ala, Gly; Asn, Asp, Glu, Gln; His, Arg, Lys; Met, Ile, Leu, Val; Phe, Tyr, Trp). (*ii*) Ties were broken by maximizing the lod score. (*iii*) No choice was made when a negative lod score resulted or when all residues were different.

Determination of Cro-like (referred to here as helix-turn-helix) DNA-binding regions followed the procedure of Dodd and Egan (2), using the parameters determined by them.

RESULTS AND DISCUSSION

DNA Database Searching for Protein Matches. The database searching procedure used a query sequence, the *Salmonella typhimurium metR* predicted protein (MetR), a recently published sequence of 276 amino acids for a gene regulatory protein (15) that was not present in available databases. Each entry in either GenBank 52 or EMBL 14 was conceptually translated into protein in the six possible reading frames, and

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

†To whom reprint requests should be addressed.
§Present address: Department of Biology, University of Saskatchewan, Saskatoon, SK, Canada S7N 0W0.

¶This sequence is being deposited in the EMBL/GenBank data base (IntelliGenetics, Mountain View, CA, and Eur. Mol. Biol. Lab., Heidelberg) (accession no. J03862).

Table 1. Best database matches to *S. typhimurium* metR activator protein

Score	Matches	Protein	Database
92	24	*E. coli* LysR activator	P, G
91	28	*E. cloacae* AmpR activator	E
91	24	*A. eutrophus* TfdO ORF	E
90	31	*S. typhimurium* LeuO ORF*	G, E
90	27	*E. coli* IlvY activator	G, E
88	31	*E. coli* LeuO ORF*	G, E
87	22	*E. coli* CysB activator	G, E
86	23	*E. coli* AntO ORF	E
85	23	*R. meliloti* NodD activator	P, G, E
85	23	*R. leguminosarum* NodD activator	G, E
85	22	*R. trifolii* NodD activator	G, E
85	21	*Bradyrhizobium* sp. NodD activator	E
85	20	*Rhizobium* sp. NodD1 activator	E
85	20	*S. typhimurium* CysB activator	G, E
85	18	Mouse ubiquitin mRNA, inverted	G
84	21	HSV-1 glycoprotein mRNA, inverted	G, E
84	18	Rabbit β-myosin mRNA, inverted	G, E
84	18	Bovine protein C mRNA, inverted	E

P, NBRF-PIR; G, GenBank; E, EMBL databases. ORF, open reading frame.
**R. meliloti* nodD protein used as probe.

each frame was compared to MetR. As each DNA database consists of ≈15 million nucleotides of sequence, translation into 6 reading frames generated 30 million amino acids, or ≈100,000 potential protein sequences of 300 residues.

The 16 best matches are shown in Table 1, along with two other matches that were found when the proteins detected in the first search were used as query proteins in similar searches. Both the lod score (10), based on comparison of all residues within the window, and the number of matches are shown. Of the 14 best matches, 10 correspond to the coding regions for bacterial activator proteins. They are *Escherichia coli* LysR, *Enterobacter cloacae* AmpR, *E. coli* IlvY, five NodD proteins from various *Rhizobia*, and the two CysB proteins from *E. coli* and *S. typhimurium*. Another is an unidentified *E. coli* protein encoded upstream of *ant*, the Na^+/H^+ antiporter gene (AntO). The other three are incomplete predicted NH_2-terminal sequences: one is found upstream of the *E. coli leu* operon (LeuO), another is its homologue in *S. typhimurium*, and another is found upstream of *Alcaligenes eutrophus tfdA* gene (TfdO).[‖] The next best matches are clearly spurious, as they detect proteins predicted from the opposite strands of animal cDNAs encoding various unrelated proteins. Excluding proteins of synonomous function, such as the NodDs, nine different proteins appear to be related to one another on the basis of database searching. Similar searches performed with the NBRF-PIR 15 database detected LysR and NodD, the only members of this group of proteins present. As LysR was the first member of this group to be studied in molecular detail, this group is referred to as the LysR family.

Sequence of LeuO. If the relationships detected above are real, then one expectation is that the 75-amino acid predicted sequence of *E. coli* LeuO present on the database is the NH_2 terminus of a LysR family member. The entire *E. coli* sequence between the *leu* operon and the *ilvIH* operon was determined (Fig. 1). Translation of this sequence extends the open reading frame from 75 to 290 amino acids. As demonstrated below, this sequence aligns throughout its length with other members of the LysR family.

[‖]The names used for putative proteins encoded by open reading frames (LeuO, AntO, and TfdO) are adopted for heuristic purpose and do not connote any relationship to *leu*, *ant*, or *tfd* genes.

```
GATTATTTCTCTGCATTCCAATAAGGGAAAGGGAGTTAAGTGTGACAGTGGAGTTAAGTA   60
                                                          M
TGCCAGAGGTACAAACAGATCATCCAGAGACGGCGGAGTTAAGCAAACCACAGCTACGCA  120
 P  E  V  Q  T  D  H  P  E  T  A  E  L  S  K  P  Q  L  R  M
TGGTCGATCTCAACTTATTAACCGTTTTCGATGCCGTGATGCAGGAGCAAAACATTACTC  180
 V  D  L  N  L  L  T  V  F  D  A  V  M  Q  E  Q  N  I  T  R
GTGCCGCTCATGTTCTGGGAATGTCGCAACCTGCGGTCAGTAACGCTGTTGCACGCCTGA  240
 A  A  H  V  L  G  M  S  Q  P  A  V  S  N  A  V  A  R  L  K
AGGTGATGTTTAATGACGAGCTTTTTGTTCGTTATGGCCGTGGTATTCAACCGACTGCTC  300
 V  M  F  N  D  E  L  F  V  R  Y  G  R  G  I  Q  P  T  A  R
GCGCATTTCAACTTTTTGGTTCAGTTCGTCAGGCATTGCAACTAGTACAAAATGAATTGC  360
 A  F  Q  L  F  G  S  V  R  Q  A  L  Q  L  V  Q  N  E  L  P
CTGGTTCAGGTTTTGAACCCGCGAGCAGTGAACGTGTATTTCATCTTTGTGTTTGCAGCC  420
 G  S  G  F  E  P  A  S  S  E  R  V  F  H  L  C  V  C  S  P
CGTTAGACAGCATTCTGACCTCGCAGATTTATAATCACATTGAGCAGATTGCGCCAAATA  480
 L  D  S  I  L  T  S  Q  I  Y  N  H  I  E  Q  I  A  P  N  I
TACATGTTATGTTCAAGTCTTCATTAAATCAGAACACTGAACATCAGCTGCGTTATCAGG  540
 H  V  M  F  K  S  S  L  N  Q  N  T  E  H  Q  L  R  Y  Q  E
AAACGGAGTTTGTGATTAGTTATGAAGACTTCCATCGTCCTGAATTTACCAGCGTACCAT  600
 T  E  F  V  I  S  Y  E  D  F  H  R  P  E  F  T  S  V  P  L
TATTTAAAGATGAAATGGTGCTGGTAGCCAGCAAAAATCATCCAACAATTAAGGGCCCGT  660
 F  K  D  E  M  V  L  V  A  S  K  N  H  P  T  I  K  G  P  L
TACTGAAACATGATGTTTATAACGAACAACATGCGGCGGTTTCGCTCGATCGTTTCGCGT  720
 L  K  H  D  V  Y  N  E  Q  H  A  A  V  S  L  D  R  F  A  S
CATTTAGTCAACCTTGGTATGACACGGTAGATAAGCAAGCCAGTATCGCGTATCAGGGCA  780
 F  S  Q  P  W  Y  D  T  V  D  K  Q  A  S  I  A  Y  Q  G  M
TGGCAATGATGAGCGTACTTAGCGTGGTGTCGCAAACGCATTTGGTCGCTATTGCGCCGC  840
 A  M  M  S  V  L  S  V  V  S  Q  T  H  L  V  A  I  A  P  R
GTTGGCTGGCTGAAGAGTTCGCTGAATCCTTAGAATTACAGGTATTACCGCTGCCGTTAA  900
 W  L  A  E  E  F  A  E  S  L  E  L  Q  V  L  P  L  P  L  K
AACAAAACAGCAGAACCTGTTATCTCTCCTAGCATGAAGCTGCCGGGCGCGATAAAGGCC  960
 Q  N  S  R  T  C  Y  L  S  *
```

FIG. 1. Sequence of the region between the *E. coli leu* and *ilvIH* operons and the predicted LysR family member. DNA sequence determinations were performed as described (16). The last base shown corresponds to position −362 of the published *ilvIH* sequence (17). A potential Shine–Dalgarno sequence is boxed. Translation is assumed to initiate at the next AUG. Amino acids are identified by the single-letter code.

Alignments. The relationship between *Rhizobium meliloti* NodD1 and LysR has been reported (18, 19). However, none of the other relationships in this group appears to have been noted (15, 20–30). Therefore, an attempt was made to verify the significance of the relationships by examining all possible protein pairs for similarities. A simple way of doing this is by a dot matrix representation, examples of which are shown in Fig. 2. Clearly, LysR is similar to MetR, CysB is similar to IlvY, and NodD is similar to LeuO, since a large number of dots are found along the diagonal with almost no off-diagonal dots for these comparisons. However, in the comparison between IlvY and AraC, a bacterial activator protein that was not detected in the searches, no dots are seen along the diagonal.

To systematically evaluate all possible comparisons, an "alignment matrix" was constructed. This consists of values representing the quality of each pairwise comparison. One such value is the number of dots that lie along the diagonal of a dot matrix plot. For example, the MetR/CysB dot matrix has 19 dots along the diagonal, whereas the NodD/LeuO dot matrix has 46. The alignment matrix (Fig. 3) tabulates dots for the nine different proteins detected in the search and for seven selected activator proteins that were not detected. Three unrelated proteins known not to have a gene regulatory function, *E. coli* aspartate transcarbamylase, horseradish peroxidase, and *E. coli* β-lactamase were also included. Nearly all proteins in the comparison are about the same size, ≈300 amino acids long.

It is apparent that there are two groups of related proteins, as shown by the triangular sections of the matrix. These groups are the LysR and OmpR protein families. The nine LysR family members show high values for nearly all combinations. CysB, one of the exceptions, appears to be

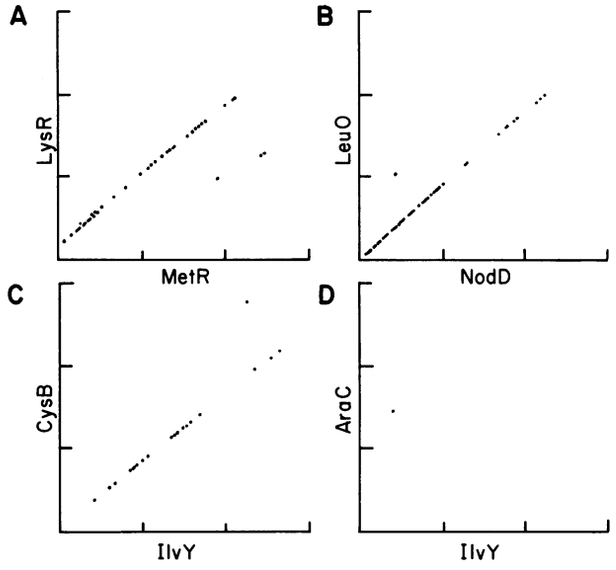

FIG. 2. Examples of dot matrix comparisons. Coordinates are in units of 100 amino acids.

only distantly related to NodD, AmpR, LeuO, and AntO. However, its strong similarities to MetR, LysR, and IlvY demonstrate that it is a member of the LysR family. Likewise, AntO seems unrelated to NodD in this analysis; however, its similarities to MetR, AmpR, LeuO, and TfdO place it within the LysR family.

The relationships among the six OmpR family members have been reported (6, 8, 32–34). It is possible that this group is related to the LysR family; however, the occasional nonzero values must be regarded as marginal and possibly spurious. The AraC protein does not appear to belong to either family.

Consensus Alignment and Search. All nine LysR family members were aligned, a procedure that required only one to six small gaps per protein (Fig. 4). Wherever four or more aligned residues were identical, this number is indicated. Although these proteins are most similar near their amino termini, short regions of strong similarity are found throughout. This indicates that these proteins are likely to resemble one another in overall properties: otherwise, such extensive similarities would not be expected. The simplest explanation is that these proteins are related by common ancestry and are similarly folded.

The multiple alignment allowed a 286-amino acid consensus sequence to be determined. When this sequence was compared to the other sequences in the alignment matrices, the scores greatly improved for members of the LysR family, but not for the others (Fig. 3, Cons column). Dot counts increased to 30–78 for this family, compared with 0–5 for the other proteins. Some of this improvement for the LysR family resulted from the contribution of each individual member to the consensus. Nevertheless, the failure of the other proteins to improve further suggests that the occasional high alignment scores outside of the LysR family were spurious.

The use of a consensus sequence for database searching can improve the likelihood of detecting family members (39). The 286-residue consensus sequence was used in searches of GenBank 52, EMBL 14, and NBRF-PIR 15 databases. Lod scores for the best stretches of each LysR family member ranged from 89 to 100, whereas spurious matches scored no better than 82 (data not shown). In spite of the greatly improved sensitivity, these searches revealed no new LysR family members. Therefore, the searching protocol used for individual sequences was sufficiently sensitive for this particular protein family by using current databases.

Assignment of a Helix-Turn-Helix DNA-Binding Motif. Others had suggested that LysR has a DNA-binding helix-turn-helix motif beginning at residue 21 (1, 2), based on sequence comparisons to known and suspected motifs. The likelihood of this motif existing within the other members of the LysR family was evaluated. The parameters of Dodd and Egan (2) were used to calculate a score that measures resemblance of any stretch of 20 residues to the helix-turn-helix DNA-binding motif. These parameters were derived by Dodd and Egan from a master set of 37 proteins that are known or are strongly suspected to have helix-turn-helix DNA-binding motifs. The 20 residues aligned with LysR amino acids 21–40 yielded scores ranging from 567 to 1778 for the family members (Table 2). To do a statistical evaluation of these values, scores were also calculated for all other possible 20 amino acid stretches of each protein. From these data, a mean and SD characteristic of each protein was calculated and used to determine the likelihood that the

	Cons	LeuO	CysB	LysR	IlvY	MetR	NodD	AmpR	AntO	TfdO	OmpR	VirG	Dye	PhoO	PhoB	NtrC	AraC	ATC	HRP
LeuO	50	-	3	14	12	11	46	10	14	14									
CysB	60	3	-	28	19	19	4	4	5	11									
LysR	78	14	28	-	29	32	18	20	6	27									
IlvY	64	12	19	29	-	32	18	15	8	31									
MetR	60	11	19	32	32	-	20	28	14	14									
NodD	61	46	4	18	18	20	-	16	0	17									
AmpR	41	10	4	20	15	28	16	-	12	22									
AntO	30	14	5	6	8	14	0	12	-	14									
TfdO	50	14	11	27	31	14	17	22	14	-									
OmpR	0	4	7	6	0	6	2	0	2	0	-	50	44	51	46	20			
VirG	0	1	4	2	0	2	12	0	0	0	50	-	34	31	18	21			
Dye	4	0	0	0	0	0	0	0	0	0	44	34	-	40	30	9			
PhoO	5	0	0	11	0	3	3	4	0	0	51	31	40	-	36	26			
PhoB	0	0	0	4	0	0	0	0	0	1	46	18	30	36	-	9			
NtrC	0	0	0	2	0	1	0	1	0	0	20	21	9	26	9	-			
AraC	0	0	0	0	0	8	0	0	1	0	0	0	0	0	0	0			
ATC	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0		
HRP	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
Blac	0	0	2	0	2	0	0	0	0	0	0	8	0	0	0	0	0	2	0

LysR family

OmpR family

FIG. 3. Alignment matrix of dot counts from the diagonals of dot matrix comparisons, examples of which are shown in Fig. 2. Comparisons between members of the same family are indicated by lines. Cons, 286-residue consensus sequence for the nine LysR family members: *E. coli* LeuO, CysB (20), LysR (21), IlvY (22), and AntO (25); *S. typhimurium* MetR (15); *R. meliloti* NodD1 (23); *E. cloacae* AmpR (24) and *A. eutrophus* TfdO (26). OmpR, *E. coli ompR* regulatory protein (31); VirG, *A. tumefaciens virG* regulatory protein (8); Dye, *E. coli dye* gene product (6); PhoO, predicted protein from an open reading frame upstream of *E. coli phoM* (32); PhoB, *E. coli phoB* regulatory protein (33); NtrC, *Klebsiella pneumoniae ntrC* regulatory protein (34). AraC, *E. coli araC* regulatory protein (35); ATC, *E. coli* aspartate transcarbamylase catalytic subunit (36); HRP, horseradish peroxidase (37); Blac, β-lactamase from Tn*3* (38).

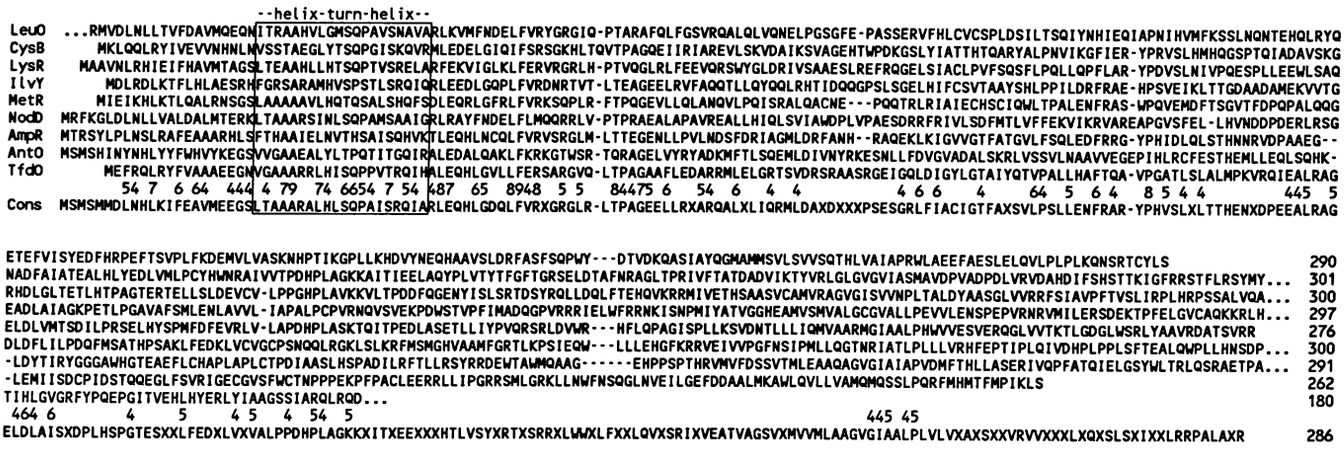

FIG. 4. Multiple alignment and consensus for LysR family proteins. Numbers below represent the number of identical residues at a position where 4 or more matches are found. Amino acids are identified by the single-letter code.

aligned region conforms to the Dodd–Egan helix-turn-helix criteria. The resulting score is seen to be highly significant in each case, between 2.6 and 5.7 SD above the mean, with the consensus sequence for this region 5.0 SD above the mean (Table 2). In nearly every case, the region aligning with LysR residues 21–40 was the highest scoring region of the protein. Only the AntO sequence included a region with a score slightly higher than that aligning with LysR residues 21–40. These results argue for the presence of a helix-turn-helix DNA-binding motif at the aligned position for all 9 proteins.

One reason for the lower scores found for several of the aligned regions than for the 20-residue segments comprising the master set (which score between 1320 and 2684) is likely to be the underrepresentation of this family in the calculation of parameters (2). LysR is the only member of this family in the master set, and it has one of the lower scores (1475). Other families contributed disproportionately to the parameters: for example, the inversion resolvase family of proteins contributed eight members to the master set. These would be expected to skew the distribution of preferred residues. The problem of skewed sampling is particularly apparent from the fact that alanine is seen at position 4 only in LysR and one other protein from the master set, whereas it is one of the most conserved residues of the LysR family, present seven of nine times. Further evidence that helix-turn-helix motifs can be found with relatively low scores is the fact that *E. coli* TrpR, known from its three-dimensional structure to have this motif at its DNA binding site (40), scores only 1026, below the minimum to be considered a member of the master set.

Table 2. Helix-turn-helix predictions for the aligned region of each protein

Protein	1st residue	Score	Mean	SD
AmpR (*E. cloacae*)	23	1778	−440	5.7
LeuO (*S. typhimurium*)	19	1343	−404	5.1
LeuO (*E. coli*)	19	1308	−363	4.7
LysR (*E. coli*)	21	1475*	−464	4.6
IlvY (*E. coli*)	18	1328	−433	4.4
CysB (*E. coli*)	19	1213	−428	4.2
NodD (*Bradyrhizobia*)	22	850	−438	3.5
MetR (*S. typhimurium*)	19	931	−412	3.3
TfdO (*A. eutrophus*)	18	906	−466	3.3
NodD2 (*R. meliloti*)	23	809	−452	3.3
NodD1 (*R. meliloti*)	26	749	−464	3.2
NodD (*R. trifolii*)	23	604	−461	2.8
ant ORF (*E. coli*)	23	647†	−388	2.7
NodD (*R. leguminosarum*)	23	567	−452	2.6
Consensus	23	1539	−435	5.0

Predictions are based on parameters from the weight matrix of 37 Cro-like binding domains (2). Mean represents average score for all possible 20-residue segments excluding that with the maximum score for each protein. Number of SD above the mean for each protein are given.

*Based on the master set excluding LysR.

†Maximum score was 668, beginning at residue 48.

It should be noted that for three of the proteins, a helix-turn-helix DNA binding motif was hypothesized to exist at other positions. However, in all three cases, the calculated Dodd–Egan scores are quite low: −329 for CysB residues 137–158 (20), +262 for NodD residues 30–49 (28), and −15 for AmpR residues 223–242 (24). It is therefore unlikely that helix-turn-helix DNA binding motifs exist at these positions. This underscores the value of identification of a protein family, particularly when one attempts to detect functional roles of peptide regions based on sequence alone.

The question of possible relationships among the different regulatory gene families can be investigated by using helix-turn-helix predictions. If the OmpR and AraC families were related to the LysR family, then helix-turn-helix predictions would be expected near the amino termini. However, in each case, helix-turn-helix predictions are made for stretches found near the middle of each OmpR and AraC family protein (data not shown). This suggests that the other known bacterial activator proteins are not homologous to members of the LysR family.

Other Features of LysR Family Members. On the one hand, members of this family have several common features: (*i*) All six proteins of known function are required for transcription of other genes. (*ii*) Five of the activator genes are each transcribed divergently from one of the operons that is regulated. (*iii*) At least four of the proteins activate amino acid biosynthetic operons. (*iv*) The levels of at least four of the proteins (LysR, CysB, NodD, and MetR) are reported to be autogenously regulated (21, 41–43). On the other hand, no consistent biochemical relationship is apparent among the hypothesized inducer molecules for any of the regulatory proteins: diaminopimelic acid for LysR (21), either acetohydroxybutyrate or acetolactate for IlvY (22), flavenoids for NodD (41), vitamin B_{12} for MetR (44), *O*-acetylserine for CysB (45), and β-lactamase for AmpR (24). Also, no obvious consensus is seen among the sequences that have been proposed as DNA binding sites for certain of the proteins (data not shown).

Little is known about the unidentified open reading frames that were detected as members of this family. In the case of LeuO, there is no evidence that it is involved in regulation of the *leu* operon upstream or the *ilvIH* operon downstream

(unpublished results). Similarly, AntO and TfdO are not known to be involved in regulation of nearby genes (refs. 25, 26, and 46; G. Mackie, personal communication).

In summary, the LysR protein family has been shown to include at least nine proteins, six of which are known bacterial activators. These relationships, initially found by DNA database searching, were confirmed by complete sequencing of a proposed family member and by pairwise and consensus comparisons. Alignment of the nine sequences allowed for a likely helix-turn-helix DNA binding domain to be identified near each amino terminus. As the methods used here required only a personal computer they should be generally useful for the routine identification of other groups of related proteins.

We thank Ed Morgan and Stanley Brown for comments on the manuscript. This work was supported in part by National Institutes of Health Grant GM29009.

1. Pabo, C. O. & Sauer, R. T. (1984) *Annu. Rev. Biochem.* **53,** 293–321.
2. Dodd, I. B. & Egan, J. B. (1987) *J. Mol. Biol.* **194,** 557–564.
3. Klug, A. & Rhodes, D. (1987) *Trends Biochem. Sci.* **12,** 464–469.
4. Wu, J., Anderton-Loviny, T., Smith, C. A. & Hartley, B. S. (1985) *EMBO J.* **4,** 1339–1344.
5. Stokes, H. W. & Hall, B. G. (1985) *Mol. Biol. Evol.* **2,** 478–483.
6. Drury, L. S. & Buxton, R. S. (1985) *J. Biol. Chem.* **260,** 4236–4242.
7. Valentin-Hansen, P., Larsen, J. E., Hjrup, P., Short, S. A. & Barbier, C. S. (1986) *Nucleic Acids Res.* **11,** 2215–2228.
8. Winans, S. C., Ebert, P. R., Stachel, S. E., Gordon, M. F. & Nester, E. W. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 8278–8282.
9. Henikoff, S. & Wallace, J. C. (1988) *Nucleic Acids Res.* **16,** 6191–6204.
10. Dayhoff, M. O., Barker, W. C. & Hunt, L. T. (1983) *Methods Enzymol.* **91,** 524–545.
11. Lipman, D. J. & Pearson, W. R. (1985) *Science* **227,** 1435–1441.
12. Maizel, J. V., Jr., & Lenk, R. P. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 7665–7669.
13. Feng, D. F. & Doolittle, R. F. (1987) *J. Mol. Evol.* **25,** 351–360.
14. Barton, G. J. & Sternberg, M. J. (1987) *J. Mol. Biol.* **198,** 327–337.
15. Plamann, L. S. & Stauffer, G. V. (1987) *J. Bacteriol.* **169,** 3932–3937.
16. Smith, D. R. & Calvo, J. M. (1980) *Nucleic Acids Res.* **8,** 2905–2919.
17. Haughn, G. W., Squires, C. H., DeFelice, M., Lago, C. T. & Calvo, J. M. (1985) *J. Bacteriol.* **163,** 186–198.
18. Horvath, B., Bachem, C. W., Schell, J. & Kondorosi, A. (1987) *EMBO J.* **6,** 841–848.
19. Appelbaum, E. R., Thompson, D. V., Idler, K. & Chartrain, N. (1988) *J. Bacteriol.* **170,** 12–20.
20. Ostrowski, J., Jagura-Burdzy, G. & Kredich, N. M. (1987) *J. Biol. Chem.* **262,** 5999–6005.
21. Stragier, P. & Patte, J.-C. (1983) *J. Mol. Biol.* **168,** 333–350.
22. Wek, R. C. & Hatfield, G. W. (1986) *J. Biol. Chem.* **261,** 2441–2450.
23. Egelhoff, T. T., Fisher, R. F., Jacobs, T. W., Mulligan, J. T. & Long, S. R. (1985) *DNA* **4,** 241–248.
24. Honore, N., Nicolas, M. H. & Cole, S. T. (1986) *EMBO J.* **5,** 3709–3714.
25. Mackie, G. A. (1986) *Nucleic Acids Res.* **14,** 6965–6981.
26. Streber, W. R. (1987) *J. Bacteriol.* **169,** 2950–2955.
27. Haughn, G. W., Wessler, S. R., Gemmill, R. M. & Calvo, J. M. (1986) *J. Bacteriol.* **166,** 1113–1117.
28. Shearman, C. A., Rossen, L., Johnston, A. W. & Downie, J. A. (1986) *EMBO J.* **5,** 647–652.
29. Schofield, P. R. & Watson, J. M. (1986) *Nucleic Acids Res.* **14,** 2891–2903.
30. Scott, K. F. (1986) *Nucleic Acids Res.* **14,** 2905–2919.
31. Wurtzel, E. T., Chou, M.-Y. & Inouye, M. (1982) *J. Biol. Chem.* **257,** 13685–13691.
32. Drury, L. S. & Buxton, R. S. (1985) *J. Biol. Chem.* **260,** 4236–4242.
33. Amemura, M., Makino, K., Shinagawa, H. & Nakata, A. (1986) *J. Bacteriol.* **168,** 294–302.
34. Drummond, M., Whitty, P. & Wootton, J. (1986) *EMBO J.* **5,** 441–447.
35. Stoner, C. M. & Schleif, R. (1982) *J. Mol. Biol.* **154,** 649–652.
36. Schachman, H. K., Pauza, C. D., Navre, M., Karels, M. J., Wu, L. & Yang, Y. R. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 115–119.
37. Welinker, K. G. (1976) *FEBS Lett.* **72,** 19–23.
38. Heffron, F., McCarthy, B. J., Ohtsubo, H. & Ohtsubo, E. (1979) *Cell* **18,** 1153–1163.
39. Patthy, L. (1987) *J. Mol. Biol.* **198,** 567–577.
40. Schevitz, R. W., Otwinowski, Z., Joachimiak, A., Lawson, C. L. & Sigler, P. B. (1985) *Nature (London)* **317,** 782–786.
41. Downie, J. A. & Johnston, A. W. (1986) *Cell* **47,** 153–154.
42. Urbanowski, M. L. & Stauffer, G. V. (1987) *J. Bacteriol.* **169,** 5841–5844.
43. Bielinska, A. & Hulanicka, D. (1986) *J. Gen. Microbiol.* **132,** 2571–2576.
44. Rowbury, R. J. (1983) in *Amino Acids: Biosynthesis and Genetic Regulation*, eds. Herrmann, K. M. & Somerville, R. L. (Addison-Wesley, Reading, MA), pp. 191–211.
45. Kredich, N. M. (1971) *J. Biol. Chem.* **246,** 3474–3484.
46. Goldberg, E. B., Arbel, T., Chen, J., Karpel, R., Mackie, G. A., Schuldiner, S. & Padan, E. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 2615–2619.