JOURNAL OF BACTERIOLOGY, Sept. 1995, p. 4872–4880 Vol. 177, No. 17
0021-9193/95/$04.00+0

A Consensus Sequence for Binding of Lrp to DNA

YUHAI CUI,[1] QING WANG,[2]† GARY D. STORMO,[3] AND JOSEPH M. CALVO[1]*

Section of Biochemistry, Molecular and Cell Biology[1] and Section of Genetics and Development,[2] Cornell University, Ithaca, New York 14853, and Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, Colorado 80309[3]

Received 6 January 1995/Accepted 20 May 1995

Lrp (leucine-responsive regulatory protein) is a major regulatory protein involved in the expression of numerous operons in *Escherichia coli*. For *ilvIH*, one of the operons positively regulated by Lrp, Lrp binds to multiple sites upstream of the transcriptional start site and activates transcription. An alignment of 12 Lrp binding sites within *ilvIH* DNA from two different organisms revealed a tentative consensus sequence AGAAT TTTATTCT (Q. Wang, M. Sacco, E. Ricca, C. T. Lago, M. DeFelice, and J. M. Calvo, Mol. Microbiol. 7:883–891, 1993). To further characterize the binding specificity of Lrp, we used a variation of the Selex procedure of C. Tuerk and L. Gold (Science 249:505–510, 1990) to identify sequences that bound Lrp out of a pool of 10^{12} different DNA molecules. We identified 63 related DNA sequences that bound Lrp and estimated their relative binding affinities for Lrp. A consensus sequence derived from analysis of these sequences, YAGHAWATTWT DCTR, where Y = C or T, H = not G, W = A or T, D = not C, and R = A or G, contains clear dyad symmetry and is very similar to the one defined earlier. To test the idea that Lrp in the presence of leucine might bind to a different subset of DNA sequences, we carried out a second selection experiment with leucine present during the binding reactions. DNA sequences selected in the presence or absence of leucine were similar, and leucine did not stimulate binding to any of the sequences that were selected in the presence of leucine. Therefore, it is unlikely that leucine changes the specificity of Lrp binding.

Evidence accumulated in the past few years indicates that Lrp (leucine-responsive regulatory protein) is an *Escherichia coli* master regulatory protein (10). Some of the operons regulated by Lrp, including *ilvIH* (30), *serA* (22, 32), *tdh* (22, 32), *sdaA* (22), *oppABCD* (1), *livJ* (18), *livKHMGF* (18), *papBA* (9), *fanABC* (9), *fim* (8), *lysU* (15, 21), and *lrp* (20, 43), were identified by genetic studies using null mutants of *lrp*. Other members of the Lrp regulon, including *ompF*, *ompC*, and *gltD*, were identified by a two-dimensional gel electrophoresis analysis (13), and yet others were identified by a p*lac*Mu transposition analysis (20).

Many of the operons that are regulated by Lrp are also subject to control by leucine, and one of the most interesting aspects of the Lrp regulon is the number of different regulatory patterns that have been observed. Lrp activates expression of some operons and represses expression of others. Among operons activated by Lrp, in some cases that activation is overcome by leucine, in other cases the activation requires leucine, and in yet other cases the activation is independent of leucine (10, 13, 20, 25). Similarly, for those operons whose expression is repressed by Lrp, the same three subcategories have been recognized; leucine overcomes the effect, leucine is required for the effect, or leucine has no effect.

The gene coding for Lrp (*lrp*) has been cloned (1, 8, 9, 22, 30) and sequenced (1, 45). A mutational analysis suggested that Lrp contains three functional domains of about equivalent sizes: an N-terminal domain containing the site for DNA binding, a middle domain responsible for transcription activation, and a C-terminal domain that is required for the response to leucine (29). Lrp has been purified to near homogeneity. It is a moderately abundant basic protein with a monomer size of 20 to 21.5 kDa as estimated by sodium dodecyl sulfate-polyacrylamide gel electrophoresis and a calculated size of 18.8 kDa (45). The native molecular mass of Lrp as measured by gel filtration is 43 kDa, suggesting that Lrp is a dimer in solution (45).

For at least the *ilvIH* operon, Lrp regulates gene expression directly by binding to specific DNA sequences (42, 44). Upon binding, Lrp bends DNA (41). In addition, Lrp binding to neighboring sites is highly cooperative and leads to the formation of a large nucleoprotein complex (40). Six Lrp binding sites have been identified by methidium propyl-EDTA-iron(II) footprinting analysis of the *ilvIH* regulatory regions of *E. coli* and *Salmonella typhimurium* (42). Comparison of a total of 12 sites from the two organisms yielded a consensus sequence, AGAATTTTATTCT, and a synthetic DNA molecule containing this sequence exhibited a relatively high affinity for Lrp (40). However, sequences similar to this consensus are not easily discerned in some other genes regulated by Lrp. For this reason, it seemed important to gain further information on a consensus sequence for Lrp binding but by using binding sites that were unrelated to those in the *ilvIH* operon.

In this study, we used a variation of the Selex strategy developed by Tuerk and Gold (39) to further study the sequence specificity of Lrp-DNA interactions. The results provided here indicate that Lrp binds in a sequence-specific manner to a consensus sequence YAGHAWATTWTDCTR, where Y = C or T, H = not G, W = A or T, D = not C, and R = A or G. This consensus is very similar to that derived from a comparison of binding sites within *ilvIH* DNA.

* Corresponding author. Mailing address: Section of Biochemistry, Molecular and Cell Biology, Cornell University, 451 Biotechnology Building, Ithaca, NY 14853. Phone: (607) 255-2437. Fax: (607) 255-2428. Electronic mail address: jmc22@Cornell.edu.

† Present address: Department of Human Genetics, University of Utah Health Sciences Center, Salt Lake City, UT 84112.

MATERIALS AND METHODS

Bacterial strains and phage. *E. coli* JM101 cells were used as the host for phage M13mp18 (New England Biolabs, Inc.) or its derivatives.

DNA synthesis. The following oligonucleotides were synthesized at the Cornell University Oligonucleotide Synthesis Facility. Underlined sequences are cleav-

age sites for *Eco*RI (oligonucleotides A, B, and D), *Bam*HI (oligonucleotides A, C, and D), *Xba*I (oligonucleotides E and F), and *Hin*dIII (oligonucleotides E and G). Sequences of the oligonucleotides are as follows: A, 5′-TGACGAATTC $ACGTGN_{20}$GTACGGATCCATGCG-3′; B, 5′-TGACGAATTCACGTG-3′; C, 5′-CGCATGGATCCGTAC-3′; D, 5′-TGACGAATTCACGTGAATGGAGG ATTTTATCGTTTGTACGGATCCATGCG-3′; E, 5′-ATGCGTCTAGAGCG TN_{20}GGACAAGCTTCGATC-3′; F, 5′-ATGCGTCTAGAGCGT-3′; and G, 5′-GATCGAAGCTTGTCC-3′. N means that either G, A, T, or C was inserted at the indicated position.

Selex-A is a mixture of double-stranded DNA fragments each 50 bp in length. Fifteen base pairs at each end are of defined sequence, but the 20 bp in the middle are random. Altogether, there are 4^{20} (10^{12}) different sequences within Selex-A. Selex-A DNA was prepared by primer extension using oligonucleotide A as the template and oligonucleotide C as the primer. Ten micrograms of template oligonucleotide and 4 μg of primer were mixed with 50 mM Tris-HCl (pH 8.0), 5 mM $MgCl_2$, 5 mM dithiothreitol, 50 mM KCl, 1.7 mM each of four deoxyribonucleotide triphosphates (dNTPs), and 30 U of reverse transcriptase (Molecular Genetic Resources) in a volume of 30 μl, and the sample was incubated at 37°C for 1 h. The primer-extended products (50 bp long) were purified by polyacrylamide gel electrophoresis, and portions were labeled with [γ-^{32}P]ATP and T4 polynucleotide kinase for subsequent gel retardation analysis. In some cases, Selex-A was cut with *Eco*RI and *Bam*HI and labeled with [α-^{32}P]dATP and reverse transcriptase, using conditions described by Wang and Calvo (40).

Selex-E was prepared by primer extension of template oligonucleotide E, using oligonucleotide G as the primer, as described above for Selex-A. In some cases, Selex-E was cut with *Xba*I and *Hin*dIII. Selex-D is a 50-bp double-stranded DNA fragment in which the 20-bp middle region is Lrp binding site 4 within *ilvIH* DNA (40). Selex-D was prepared by primer extension using oligonucleotide D as the template and oligonucleotide C as the primer. In some cases, Selex-D was cut with *Eco*RI and *Bam*HI.

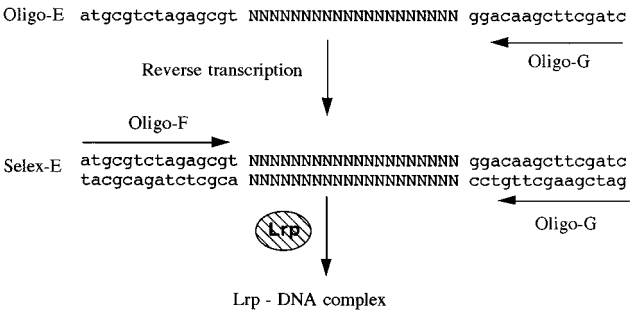
Selection of DNA fragments that bind Lrp. ^{32}P-end-labeled DNA was mixed with purified Lrp (45) in a 200-μl volume containing 20 mM Tris-HCl (pH 8), 0.4 mM EDTA (pH 8), 0.1 mM dithiothreitol, 50 mM NaCl, 1 mM $MgCl_2$, 12.5% glycerol, and 100 ng of bovine serum albumin per ml. For the first cycle of selection, 750 nM Selex-E DNA (5 μg) and 13.3 nM Lrp (dimer) were used. Reaction mixtures were incubated for 20 min at room temperature, and then samples were applied to three 1.2-cm wells of a 1.5-mm-thick 8% polyacrylamide gel (acrylamide/bisacrylamide, 40:1) and fractionated at 10 V/cm at room temperature. The wet gels were wrapped with Saran Wrap and exposed to X-ray film. To recover DNA, samples of gel were excised, soaked overnight at 37°C in 3 ml of elution buffer (0.5 M ammonium acetate, 10 mM magnesium acetate, 1 mM EDTA), and precipitated with ethanol. After the sample was dissolved in 50 μl of TE buffer (10 mM Tris [pH 8.0], 1 mM EDTA), a 5-μl sample was added to a reaction mixture together with 0.2 μM oligonucleotide F, 0.2 μM oligonucleotide G, and 5 U of *Taq* DNA polymerase in a final volume of 100 μl containing 50 mM KCl, 10 mM Tris-HCl (pH 8.8), 1 mM $MgCl_2$, 0.01% gelatin, and 0.5 mM each dNTP. Six such reaction mixtures were set up. All of the components except the template and enzyme were maintained at 75°C for 5 min before addition of the template and enzyme, and then the samples were heated at 94°C for 2 min (hot start). For each of 12 cycles of PCR, samples were denatured at 92°C for 1 min, annealed at 46°C for 1 min, and extended at 72°C for 1 min. All six reaction mixtures were pooled, and the DNA was isolated by ethanol precipitation. The concentration of DNA was determined by electrophoresis, and then about 10% of the sample was labeled with ^{32}P with the aid of polynucleotide kinase. About half of the DNA was used for the next cycle of the Selex procedure.

The second and third cycles of the Selex procedure were performed as described above except that 120 nM (800 ng) of DNA and 6.7 nM Lrp were used.

The conditions described above were worked out in a preliminary experiment that employed Selex-A as a template, primers B and C, and PCR conditions that used as many as 30 cycles. Ten cycles of the Selex procedure were required to select sequences that bound strongly to Lrp. The following was learned in an attempt to understand why so many cycles were required. The number of PCR cycles is critical: in our experiments, the amount of product markedly diminished as the number of cycles increased beyond 15. One possible reason for this is that as the primer disappears, annealing of strands after denaturation leads to products that are double stranded at each end (unique sequences) and single stranded in the middle 20 bp. Indeed, the decrease in the amount of 50-mer product with increasing cycle number was correlated with an increase in the amount of a more slowly migrating substance that was heterogeneous in size. Another problem came to light with a control PCR experiment that yielded an apparent 50-mer product when template was omitted from the reaction. This problem was minimized by reducing the number of PCR cycles to 12 and by using hot starts.

Selection of DNA fragments that bind Lrp in the presence of leucine. The Selex method using Selex-E as the template and primers F and G was carried out in the presence of 7.5 mM leucine, using the procedures described above. The two experiments were actually carried out at the same time, to minimize any problems of template contamination of PCRs.

Cloning and sequencing the products of selection. DNA amplified from the last cycle of selection was cut with *Xba*I and *Hin*dIII (*Eco*RI and *Bam*HI for the preliminary experiment) and cloned into double-stranded phage M13mp18 DNA cut with the same enzymes. After transformation and plating, individual plaques were purified and single-stranded DNA prepared from them was sequenced by the dideoxy method of Sanger et al. (35), using the Sequenase kit from United States Biochemical, Inc.



1. Isolate Lrp - DNA complex by gel electrophoresis
2. Amplify DNA by PCR with primers F and G
3. Repeat cycle of binding, electrophoresis, and PCR multiple times
4. Clone and sequence members of the population

FIG. 1. Strategy for selecting DNA sequences that bind Lrp.

Analysis of Lrp binding to specific sites. Double-stranded replicative-form DNA was prepared from cells infected with the M13mp18 phage described above. After cutting with *Xba*I and *Hin*dIII (*Eco*RI and *Bam*HI for the preliminary experiment), fragments were end labeled by treatment with reverse transcriptase and an α-^{32}P-dNTP and purified by gel electrophoresis. Binding site 2 from the *ilvIH* operon, chemically synthesized previously (40), was labeled with [^{32}P]dCTP and reverse transcriptase. Labeled DNA was mixed with Lrp, and gel retardation analyses were performed as described above for selection of DNA fragments that bind Lrp.

For quantitative measurements of complex formation, gels were transferred to Whatmann 3MM paper and dried, and counts in bands containing free DNA (F) and Lrp-DNA complex (C) were measured with a Betascope model 603 blot analyzer (Betagen Corporation); the percentage of DNA in a complex was equal to C/F + C.

Other procedures. Standard DNA techniques, including digestion of DNA with restriction enzymes, labeling of DNA by T4 polynucleotide kinase or reverse transcriptase, and DNA ligations, were as described by Sambrook et al. (34) and Ausubel et al. (2). The Pileup, Prettyplot, and Consensus programs from the Genetics Computer Group (GCG) software package were used to align DNA sequences (12). Alignments were also performed by using the neural network approach of Heumann et al. (19). Quantitative binding data were analyzed as described by Stormo et al. (36) and Barrick et al. (3).

RESULTS

Selection of DNA molecules that bind Lrp. To identify DNA sequences that are recognized by Lrp, we used a variation of a protocol developed by Tuerk and Gold (39). We synthesized a 50-base oligonucleotide, oligonucleotide E, having a random sequence within the middle 20 bases (Fig. 1). Oligonucleotide G, an oligonucleotide complementary to one of the defined flanking sequences, was used as a primer to convert oligonucleotide E to a double-stranded form. The resulting mixture, expected to contain 10^{12} (4^{20}) different 50-mer oligonucleotides, is called Selex-E. Selex-D is a double-stranded 50-mer having a central 20-bp sequence that is a known Lrp binding site (binding site 4 upstream of the *ilvIH* operon) (40). Five micrograms of Selex-E (contains about 100 molecules of each of 10^{12} types) and separately about 2 ng of Selex-D, each labeled at their ends with ^{32}P, were incubated with purified Lrp and fractionated through an 8% polyacrylamide gel. Selex-D was used as a control to define the position to which an Lrp-DNA complex migrated in the gel. The corresponding region of the gel containing the Selex-E–Lrp complex was cut out, and DNA was eluted from the gel. DNA isolated in this way was amplified by PCR, using as primers oligonucleotides F and G

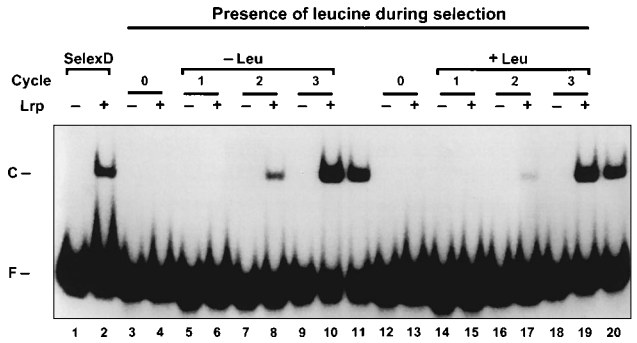


FIG. 2. Gel retardation analysis used to monitor the progress of enrichment for Lrp binding sites. About 0.1 ng each of ^{32}P-labeled Selex-D, Selex-E, and the DNAs isolated after the first, second, and third cycles of selection were incubated with (+) or without (−) 16.5 nM purified Lrp and fractionated through 8% polyacrylamide. Cycle 0 is Selex-E DNA before selection. For lanes 5 to 11, selections were carried out in the absence of leucine, whereas for lanes 14 to 19, 7.5 mM leucine was present during the selection. Lane 11 is the same as lane 10 except that the binding reaction for gel retardation contained 7.5 mM leucine (similarly for lanes 20 and 19). C, complex; F, free DNA.

(Fig. 1), and the product was subjected to the next cycle of selection that included binding, electrophoresis, elution, and amplification.

After a number of cycles, we assessed the progress of our enrichment for DNA molecules recognized by Lrp. A small fraction of the original Selex-E, and separately of each amplified product, was cut with *Xba*I and *Hin*dIII, labeled with ^{32}P, and analyzed by gel retardation using purified Lrp. For the original Selex-E and for the product of the first cycle, no Lrp-DNA complex was visible at an Lrp concentration of 16.5 nM (Fig. 2, lanes 3 to 6). For the products from cycles 2 and 3, 16.5 nM Lrp caused 7.4 and 44% of the total DNA to be in a complex, respectively. By comparison, Selex-D, containing a known weak binding site, gave 6.9% complex under the same conditions.

Sequence analysis of selected Lrp binding sites. DNA from cycle 3 was cut with *Xba*I and *Hin*dIII (sites within the defined flanking regions), and the products were ligated into the replicative form of phage M13mp18 DNA that was cut with the same enzymes. After transformation and plating, plaques were purified and replicative-form DNAs isolated from them were tested for the ability to bind Lrp. About 90% of them did bind Lrp. The nucleotide sequences of 30 such isolates and their best alignment as determined by the GCG Pileup program are shown in Fig. 3. This program produces a consensus sequence showing the nucleotide most often found at each position of the consensus. A clear consensus of at least 15 bp emerged from this analysis, with strong indications of palindromy at positions 2 and 14, 3 and 13, and 5 and 11 (Fig. 3; Table 1). Note that in performing an alignment, one of the two possible orientations of each sequence must be chosen (i.e., the top strand or the bottom strand, each read 5′ to 3′). For the alignment shown in Fig. 3, the strand that had a T at position 8 within a presumed 15-bp consensus was chosen. The rationale for this choice is as follows. Position 8 within a presumed 15-bp consensus is always an AT base pair, both for the sequences shown in Fig. 3 and for other sequences described below. Inspection of the sequences in Fig. 3 shows that the middle five positions of the consensus are generally AT rich and that there is a marked asymmetry in their distribution, with A's tending to be on one strand and T's tending to be on the other. For example, when just the three central positions 7, 8, and 9, are considered, 15 of the 30 sequences have AAA on one strand and TTT on the other. Seven other sequences have AT base pairs at the three central positions, but none of them are of the type TAT or ATA, suggesting that a clustering of A's or T's on one strand has been selected. Similarly, all eight of the remaining sequences have two AT base pairs and one GC base pair, and all of them are of the type TTC, GTT, CTT, or TTC, again suggesting clustering of AT base pairs. Thus, choosing strands for alignment that have a T at position 8 is consistent with the asymmetry observed by inspection.

Prior to obtaining the results shown in Fig. 3, we performed a preliminary experiment using a different template and set of primers. As described in Materials and Methods, the conditions for this Selex experiment were not optimal and 10 cycles were required for selection of sequences that bound Lrp tightly. Nevertheless, as shown in Fig. 4 and Table 2, the consensus derived from this preliminary experiment was very similar to that shown in Table 1 (10 of 13 positions are the same). The fact that two different sets of templates and primers gave nearly the same consensus supports the general validity of this approach for determining a consensus sequence. The suboptimal conditions of the preliminary experiment may have favored one or more selective pressures other than for Lrp binding, and that may account for the observed differences in the results of the two experiments.

Selection for Lrp binding in the presence of leucine. In vitro, leucine was found to decrease the extent of Lrp binding to multiple sites upstream of the *ilvIH* promoter (33). However, at certain concentrations of Lrp and leucine, the relative amounts of some DNA-Lrp complexes seemed to be enhanced by leucine (41a). Also, for some operons controlled by Lrp, the effect of Lrp in vivo is overcome by exogenous leucine, whereas in others, the effect of Lrp is seen only in the presence of exogenous leucine (18, 25). One possibility consistent with these results is that Lrp recognizes one consensus sequence and that Lrp with bound leucine recognizes another. To test

```
C-26    t A T G A A A A G G A T A T T A T G C T A g g a c a a
C-55    g t A G A A C A G G A T A T T A T T C T G G G g a c a
C-25    a g a g c g t A G A G A G A G T T A T G C T T G G T G A A G
C-9     g c g t A G G A G A G A G T T A T G C T G G A T g g a
C-40    c g t G A A C A G A G G T T T G T G C T G G A G g a c
C-63    g t C A G A A A G T G A A T T A T G C T G G g g a c a
C-2     g t T G G G T A G T G G T T T A T T C C A G G g a c a
13      t g t c c C T A G T G A T T T A C T G C T G C T G T a c
42      t t g t c c T G G T G T T T T G T G C C T A T T G T G T G a
52      c t t g t c c A G C C T T T T A T C C T G T G T T T A a
62      t c c T T G C A G C A T T T T A T C C C C G T g g a c g
C-39    g t A C A G G A G C A G A T T A T C C T G T g g a c a
10      G A G A T C C C A G C C A G T T T A T T C C T a c g c t c t
6       g t c c T A C G A T A T T T C C T T C C T A T T G A c g
C-61    g c g t T A C A G T A T T T C A T G C C T A G T G g g a
1       c c A C T C C C A A A A T A T T C C T G C C T A T A c g c t
59      t t g t c c c C A G T A T A T T C T C C C T T T A T A c g
49      c c C T T C T C T A G T A T C C T T A T G C C T C T A c g c t
46      A T A A C T C T C G G G A T T T T G T T C C T A c g c t c t
72      g t c c C A C C G C A T T T T T G T T C C T A C A c g c t
20      c c G T C C C C A G G A T T T C C T A C T C C C A c g c t
60      c c C T C C T C A G T A T T T T C T T C C T C C C A c g c t
33      c c G A C A T A T A G C A T T T C C T G C C T C C T A c g c t
68      t g t c c G C C G A A T T T T C T A C T G C C C C T A c
69      t t g t c c C A G C A T T T T T C T A C C T A T C C C A C A
17      g t c c T T T T A G A A C C T T T A T A C C T A T T C A C g
29      t g t c c G T A G G C C T T T T C T A C C T A T T C C A C
37      g t c c C G C A G A A C C T T C C C T A C C T G T T T A C g
67      c t t g t c c A G A G T T T T A T G C C T C T C C C C C
7       G C A C A C C C A T A T T T T T C A C C T A c g c t c t

Consensus           C A G - A T T T T - T - C T
```

FIG. 3. Alignment of sequences from 30 Lrp binding sites. The alignment was carried out with the GCG Pileup program, using a plurality value of 16 (designation as a consensus nucleotide requires at least 16 of 30 matches). One strand from each isolate was analyzed (identifier number shown on the left), chosen on the basis of having a T at position 8 of the consensus. Isolates without a C prefix have the orientation 5′...*Hin*dIII...N_{20}...*Xba*I...3′, whereas those with a C prefix have the orientation 5′...*Xba*I...N_{20}...*Hin*dIII...3′. Capital letters denote nucleotides that were within the set of 20 randomized nucleotides, and lowercase letters denote unique flanking sequences that were from the primers (Fig. 1).

TABLE 1. Summary of alignment of selected strands from each of 30 sequences[a]

Consensus[b]			C	A	G		A	T	T	T	T	a/c	T		C	T	g/a		
Position[c]			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
G%	7	27	7	10	**90**	17	27	13	7	0	0	13	0	40	0	0	37	27	33
A%	20	27	7	**80**	10	27	**67**	13	20	0	3	47	0	23	0	0	43	0	30
T%	30	10	26	0	0	33	0	**67**	**70**	**100**	**80**	3	**97**	23	0	**93**	3	47	27
C%	43	37	**60**	10	0	23	7	7	3	0	17	37	3	13	**100**	7	17	27	10

[a] The GCG Pileup program was used to produce an alignment of the sequences shown in Fig. 3, and then the Consensus program was used to generate the numbers shown.

[b] A consensus nucleotide is in capital letters if it amounts to 50% or more of the nucleotides at that position (boldface numbers). Consensus nucleotides in lowercase together sum to 80% or more.

[c] At positions 2 through 14, the alignment was such that all 30 nucleotides analyzed were derived from originally randomized sequences (Fig. 3). This was not the case for the remaining positions: position −2, 21 nucleotides present in the alignment; position −1, 24; position 1, 27; position 16, 27, position 17, 26. For these latter cases, appropriate nucleotides from adjacent unique sequences were added (lowercase letters in Fig. 3) so that in each case, 30 sequences are compared.

this possibility, the Selex procedure summarized in Fig. 1 was carried out in the presence of leucine. We chose a leucine concentration (7.5 mM) that reduced Lrp binding to *ilvIH* site 2 fourfold at an Lrp concentration that was equal to the dissociation constant. The proportions of the total DNA that formed a complex with 16.5 nM Lrp were <1, 2.6, and 25% for cycles 1, 2, and 3, respectively (Fig. 2, lanes 14 to 19). DNA from cycle 3 was cloned into vector M13mp18. Of 53 isolates tested, only 23% (12 clones) showed measurable binding to Lrp, compared with 90% of the isolates from the Selex experiment performed in the absence of leucine. The nucleotide sequences of these 12 clones are shown in Fig. 5. The consensus derived from their alignment (Table 3) is almost identical to that derived from the Selex experiment performed in the absence of leucine (Table 1).

A gel retardation analysis was performed with heterogeneous DNA from cycle 3 with binding reactions that either lacked or contained leucine. For both selection experiments (i.e., selection in the absence and presence of leucine), leucine in the binding reaction reduced binding of DNA to Lrp (Fig. 2, lanes 10 and 11, 44% complex versus 28%; lanes 19 and 20, 25% complex versus 19%). Similar experiments were performed with each of the 53 clones isolated from the Selex experiment performed with leucine. Leucine in the binding reaction reduced Lrp binding to each of the 12 clones which bound Lrp and did not cause binding of Lrp to the other 41 clones (data not shown).

Thus, conditions of selection that would favor DNA sequences that bind more tightly to Lrp in the presence of leucine yielded sequences that closely resembled those that were selected in the absence of leucine, and none of the selected sequences bound more tightly to Lrp in the presence of leucine. Given these results, it seems unlikely that the binding specificity of Lrp changes upon binding leucine.

Analysis of individual Lrp binding sites. Experiments were performed with all but four of the DNAs selected in the three experiments described above (sequences shown in Fig. 3 to 5) to determine their relative affinities for Lrp. In some cases, apparent dissociation constants were determined from titrations by using a range of Lrp concentrations and gel retardation as a measure of complex formation. Protein was in excess over DNA in these experiments, and the apparent dissociation constant was taken as the concentration of Lrp that drove half of the DNA into complex. Examples of such experiments are shown in Fig. 6. The apparent dissociation constants ranged from values similar to that of site 4 (a weak naturally occurring Lrp binding site) to values similar to that for site 2 (a strong natural site). For the remaining cases, estimates of relative affinities were made by gel retardation using a single, relatively low concentration of Lrp (10 nM monomer). The relative affinities of the 63 isolates varied over a 285-fold range. A summary of all of the data, expressed as the natural logarithm, is shown in Table 4.

Quantitative analysis of the relationship between nucleotide sequence and binding activity. The degree of similarity of selected sequences to the consensus was estimated by procedures described by Goodrich et al. (17). A similarity score was calculated for each of the 65 sequences listed in Table 4 (63 sequences analyzed here plus two natural sites). The similarity score is related to the sum derived by adding the relevant values from Table 1 at each position of the consensus (see footnote *c* to Table 4 for the exact equation used). The average similarity score for the 65 sequences is 73 ± 11. That value can be compared with an average score of 4 ± 17 for 30 random sequences having the same composition. Thus, the average similarity score for one group of the selected isolates is more than 4 standard deviations higher than for an average random sequence. Given this fact, it was surprising to find that there was only a very poor correlation between the logarithm of the relative binding and the similarity score (data not shown; correlation coefficient, $r = 0.385$). For the data assembled by Goodrich et al. for integration host factor, the correlation between the logarithm of the relative binding and the similarity score was 0.94 (17).

```
10-12      A G G A C C A G T G T T T T C G T C T A c a c g t
10-23      C A C G G C G G A A C T T T C G A C T A c a c g t
10-20      g t a c A C A G G G G A T T C T G C T G T G A T c
10-27      g t a c A C A G G G G A T T C T G C T G G G T G c
C-11-39    a c g t g C A G A G G A T T C G T C T G T C T C C
C-11-50    a c g t g T A G A G G A T T C G T C T G T C T C C
C-11-45    a c g t g C A G A G G A T T C G C C T G T T C C T
C-11-69    a c g t g T A G A G G A T T C G C C T G T T C C T
C-10-22    a c g t g C A G A G T A T T C C A C T G G T C C T
C-11-58    g t g T A C A G A G G A T T C G C C T G G A T g t
C-11-48    a c g t g C A G T C G A T T C G T C T C C G T G C
C-11-104   a c g t g C A G T G A A T C G T G C T G T C T C C
11-114     G G A G A C A G G C G T T T G T G C T G c a c g t
11-41      G G A G G C A G G C G A T T G T G C T A c a c g t
10-9       a c A T C C A G G C G T T T G T T C T G T T c a c
10-18      a c G G C C A G C A G T T T A T C C T G T A c a c
10-3       a c A G G C A G C A T A T A G T T C T G T C c a c
10-21      C C A C G G A G C A C A T T C C T C T A c a c g t
11-103     C C A C G G A G A C G A T T C T A C T G c a c g t
11-47      a c C C A C G G G A C A T T A T T C T G T A c a c
11-55      a c C C A C G G G A C A T T G T T C G G T A c a c
11-49      a c A C A C A G A A C A T T C A C C T G T A c a c
C-10-26    g C A C A C A G T A C A T T C C T C C G T g t a c
C-11-73    g A A C A C G G G C A T T T A T C C T G T g t a c
11-51      g t a c G C G G G A C A T T A T G T T G T C T T c

Consensus            C A G - - G A T T C T - C T G T
```

FIG. 4. Alignment of sequences from 25 Lrp binding sites isolated in a preliminary experiment. Details are as for Fig. 3 except that the plurality was 13 and isolates without a C prefix have the orientation 5′...*Bam*HI...N_{20}...*Eco*RI...3′, whereas those with a C prefix have the orientation 5′...*Eco*RI...N_{20}...*Bam*HI...3′.

TABLE 2. Summary of alignment of 25 sequences obtained in a preliminary experiment[a]

Consensus[b]			C	A	G			G	A	T	T	C	T		C	T	G	T	
Position[c]			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
G%	20	52	8	20	**100**	36	40	**52**	0	0	0	24	32	24	0	4	**80**	12	20
A%	4	36	0	**80**	0	36	36	8	**76**	0	4	16	4	12	0	0	16	0	44
T%	36	0	8	0	0	16	0	12	24	**100**	**92**	0	**52**	40	4	**92**	0	**60**	16
C%	40	12	**84**	0	0	12	24	28	0	0	4	**60**	12	24	**96**	4	4	28	20

[a] The GCG Pileup program was used to produce an alignment of the sequences shown in Fig. 4, and then the Consensus program was used to generate the numbers shown.

[b] A consensus nucleotide is in capital letters if it amounts to 50% or more of the nucleotides at that position (boldface numbers). Consensus nucleotides in lowercase together sum to 80% or more.

[c] At positions 1 through 15, the alignment was such that all 25 nucleotides analyzed were derived from originally randomized sequences (Fig. 4). This was not the case for the remaining positions: position −2, 16 nucleotides present in the alignment; position −1, 19; position 16, 19, position 17, 16. For these latter cases, appropriate nucleotides from adjacent unique sequences were added (lowercase letters in Fig. 4) so that in each case, 25 sequences are compared.

We also analyzed the complete set of 63 selected fragments which were shown to bind to Lrp by using the neural network multiple alignment method of Heumann et al. (19). Initially we did not assume that the sites were symmetric but merely assumed that there was a binding site in common on each fragment that could occur on either strand. Site sizes between 14 and 20 bases were tested. The results of this initial analysis were consistent with those from the Pileup analyses on the separate data sets. The most significant site size was 15 bases, and the sites appeared to be approximately symmetric. We then found the best symmetric solution of size 15 by setting weights to be identical for the complementary bases at positions 1 and 15, 2 and 14, etc. In the neural network model of Heumann et al., weights correspond to position-specific binding energies. At the center position, 8, weights were set to be identical for complementary bases. The best alignment of this type was identical to that shown in Fig. 3 to 5 except for isolate Leu-5 (see the legend to Fig. 5).

Assuming that the aligned sites are the correct binding sites, we can solve for the binding energy contributions of each base at each position which gives the best fit to the binding data by multiple regression, as in the analyses of Stormo et al. (36) and Barrick et al. (3). This method assumes that the contribution of each base pair to binding is independent of other base pairs, that is, that the total binding energy, which is proportional to the logarithm of the binding constant, is the sum of individual contributions. In an initial analysis, we assumed perfect symmetry of the binding region (for example, positions 1 and 15 were assigned the same binding energy). The matrix of binding energies derived from this analysis is shown in Fig. 7A, and the binding energies for the 63 Lrp binding sites computed from this matrix are given in Table 4 (columns labeled Stormo). A comparison of predicted versus experimental values (Fig. 7B) gave a correlation coefficient of 0.64 (corrected for degrees of freedom). The standard deviation of the difference between the observed and predicted values was 0.73, indicating that most of the binding activities were predicted within a factor of about 2.

In a second analysis, we questioned the assumption of perfect symmetry for the binding region. It seems clear that strong elements of symmetry exist within the consensus, for example, at positions 2 and 14, 3 and 13, and 5 and 11. It seemed less clear that the symmetry extends to positions 7 and 9, and for this reason we performed another multiple regression analysis assuming symmetry for positions 1 to 6 and 10 to 15 but not for positions 7 to 9. This analysis did not improve the correlation coefficient, which remained 0.64 after correcting for degrees of freedom. Finally, as a control, we also randomly assigned each of the binding affinities to one of the sequences and performed the multiple regression analysis. As expected, the corrected correlation coefficient was 0.0.

Given all of these considerations, the best consensus that emerges from this study is that derived from the quantitative matrix analysis shown in Fig. 7A: YAGHAWATTWTDCTR, where Y = C or T, H = not G, W = A or T, D = not C, and R = A or G.

DISCUSSION

The identification of a consensus sequence for Lrp is an important step in its characterization. In this study, we used the Selex strategy devised by Tuerk and Gold (39). This strategy, and similar ones devised by others (7, 23, 31, 38), was used successfully to identify the recognition sites of a number of DNA-binding proteins, including c-Myc (6), Evi-1 (28), the T/E1A-binding domain of the retinoblastoma gene product (11), Sp1 (38), and serum response factor and Fos (31). In our studies, we used a 20-bp degenerate sequence to search for DNA sequences that were recognized by Lrp. In total, 67 sequences were isolated from three separate experiments. Given that the three experiments yielded similar consensus sequences, and especially in view of the fact that one of the experiments used different primers and template, we chose to analyze the combined results of the three experiments.

We measured the relatively binding affinities of the 63 sequences and correlated these experimental values with parameters that reflected the degree of match to the consensus. If one computes similarity scores as done by Goodrich et al. (17) or takes the binding energies predicted from the neural network multiple alignment method (which does not include the quantitative binding affinities) and compares them with the measured relative binding strengths, the correlation is only

```
Leu-2      C C C C A A C A G T G C T T T A T G C T a c g c t
Leu-3      C T C G A A C A G T C A T T T C T G C T a c g c t
C-Leu-28   t A G A A C C A G T G A A T T A T G C C G g g a c
Leu-5      c A A T A G C A G T T T A T T C T C C C G a c g c
Leu-41     C A G T A C C A G C A T C T T C T G C T a c g c t
Leu-1      c c A A A C C A G T C G T T T C C A C T A T a c g
Leu-24     g t c c A T C A G A T G T T T C T C C T G C C T a
C-Leu-13   g t G G C T T A G G A T A T T C T G C T G T g g a
Leu-16     t g t c c T C A G C A T A T T C T C C C T G T A G C
Leu-31     t c c T C A A A G C A T T T T G T A C T G T C a c
C-Leu-19   c g t A A G G A G A A T A T T A T G C T A T G g g
C-Leu-53   g t A G G A C A G C A C A T T A T A C C G C g g a

Consensus          A   C A G - A T A T T C T G C T G
```

FIG. 5. Alignment of sequences isolated from the Selex experiment carried out in the presence of leucine. Details are as for Fig. 3 except that the plurality was 6. The sequence of Leu-5 is not aligned correctly, probably because of the relatively small number of sequences in this alignment. The 15 bases within Leu-5 that most closely match overall consensus derived from 63 sequences is TAGCAGTTTATTCTC.

TABLE 3. Summary of alignment of 12 sequences selected in the presence of leucine[a]

Consensus[b]			C	A	G		A	T	A	T	T	C	T	G	C	T	G		
Position[c]			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
G%	8	17	8	0	**100**	8	17	17	0	0	0	8	0	**50**	0	0	**58**	8	58
A%	67	33	8	**100**	0	17	**50**	17	**50**	0	0	33	0	25	0	0	42	8	17
T%	0	25	8	0	0	42	17	**50**	42	**100**	**100**	0	**92**	0	0	**75**	0	42	0
C%	25	25	**75**	0	0	33	17	17	8	0	0	**58**	8	25	**100**	25	0	42	25

[a] The GCG Pileup program was used to produce an alignment of the sequences shown in Fig. 5, and then the Consensus program was used to generate the numbers shown.

[b] A consensus nucleotide is in capital letters if it amounts to 50% or more of the nucleotides at that position (boldface numbers).

[c] At positions −1 through 14, the alignment was such that all 12 nucleotides analyzed were derived from originally randomized sequences (Fig. 5). This was not the case for the remaining positions: position −2, 11 nucleotides present in the alignment; position 15, 9; position 16, 7, position 17, 4. For these latter cases, appropriate nucleotides from adjacent unique sequences were added (lowercase letters in Fig. 5) so that in each case, 12 sequences are compared.

about 0.39. This is considerably lower than has been observed for similar comparisons, for example, integration host factor (17), LexA (4), and RNA polymerase (27). The poor correlation for the data described here seems surprising since these fragments were selected in vitro on the basis of their affinities to Lrp. Even the best-fit predictions of binding energy give a correlation of only 0.64, which is surprisingly low. Some possible reasons for the poor correlation follow. (i) The experimentally determined binding energies may be accurate only to within about a factor of 2. However, if that were the only cause of the variation, we would still expect to see a correlation coefficient higher than 0.64. (ii) Our analysis is based on a 15-bp consensus. While these 15 bp certainly play a significant role in binding, there may be contributions from flanking base pairs that are important for specificity that we ignored. It is important to note that the selection was performed with 50-mer oligonucleotides but the binding studies were carried out with 35-mers derived from them after cloning. It is conceivable that some extra binding strength derives from the longer sequence and that the extra amount depends on the 15-bp sequence that is related to the consensus. (iii) The models underlying most matrix analyses assume that each position makes a contribution to binding that is independent of the others, i.e., that the binding energies are additive over all positions. This assumption seems to be at least mostly true for DNA-binding proteins such as Cro (37), integration host factor (17), LexA (4), Crp (5), and RNA polymerase (27), but it may not be true for LacI (24) and perhaps not for Lrp. One potential cause of nonadditivity, flexibility, might be imagined to be important in the central AT-rich region of the consensus. However, were this the case, we might have expected the nonsymmetric multiple regression analysis to give a higher correlation, which it did not. (iv) Perhaps there is some type of selection occurring during the Selex procedure other than for binding to Lrp. For example, it is conceivable that certain sequences are better amplified during PCR than other sequences. This explanation seems unlikely in view of the fact that very similar consensus sequences were derived from experiments using two different sets of templates and primers. (v) Some of the fragments analyzed may have more than one good binding site for the protein. If a fragment has two binding sites, which are overlapping so that only one can be occupied at any time, that will essentially double the on rate of the protein-DNA interaction. (This assumes a random collision model of the on rate.) However the off rate will be the same as it is determined only by the energy of the site which is bound, not by the other potential sites on the same fragment. The real effect of this on the analysis that we did is that the binding site that we have identified by the alignments may only contribute to part of the total binding energy of the fragment. That is, if we were to measure binding affinity to that site alone, we might find that it is significantly lower than to the entire fragment. In fact, several of the fragments have multiple relatively good binding sites within them, including the one which is the largest outlier in the comparison between the predicted and observed binding energies. Fragments with multiple binding sites should be selected during the in vitro selection for the same reasons.

From the results of the quantitative analysis shown in Fig. 7A, the consensus may be formulated as YAGHAWATTWT DCTR, where Y = C or T, H = not G, W = A or T, D = not C, and R = A or G. Previously, by aligning six Lrp binding sites from *E. coli ilvIH* DNA and six sites from *S. typhimurium ilvIH* DNA, we derived a tentative consensus sequence for Lrp binding: AgaATTTTATtcT (42). To facilitate comparison, the two consensus sequences are juxtaposed in the first two lines in Table 5. The two are closely related, with clear evidence of symmetry at positions 2 and 14, 3 and 13, and 5 and 11. The major insights derived from this analysis, besides supporting the consensus derived from comparing natural sites, are that the consensus is 15 rather than 13 bp in length and that G's and C's are not favored at positions 4 and 12, respectively.

Also shown in the third line of the comparison in Table 5 is a consensus sequence derived by Rex et al. (32) from an analysis of sequences near promoters known to be affected by Lrp. This latter consensus sequence is similar to half of the palin-

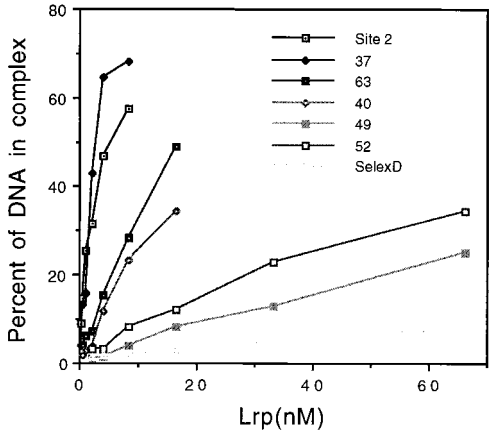


FIG. 6. Binding curves for several Lrp binding sites. Sites 2 and 4 (Selex-D) are Lrp binding sites within *ilvIH* DNA (40), whereas clones 37, 63, 40, 49, and 52 were isolated in this study. Purified DNA was incubated with purified Lrp, and complexes were separated by electrophoresis through 8% acrylamide. Radioactivity in the Lrp-DNA complex and free DNA was quantified with a Betascope blot analyzer, and the percentage of DNA in the Lrp-DNA complex was calculated. The concentration of Lrp shown is that calculated for the monomer.

TABLE 4. Summary of binding data for 63 sequences[a]

Selection in the absence of leucine				Preliminary expt				Selection in the presence of leucine			
Clone	ln relative binding strength[b]	Predicted binding strength		Clone	ln relative binding strength	Predicted binding strength		Clone	ln relative binding strength	Predicted binding strength	
		Goodrich[c]	Stormo[d]			Goodrich	Stormo			Goodrich	Stormo
55	5.193	87	4.595	10-3	5.081	72	4.824	Leu-19	5.991	84	4.633
37	5.050	77	3.481	10-12	4.804	76	3.443	Leu-16	4.875	85	4.390
61	4.820	91	4.485	10-18	4.754	86	3.856	Leu-13	4.564	83	4.530
68	4.682	84	4.321	11-47	4.127	68	2.997	Leu-2	4.174	82	4.176
69	4.605	95	4.957	11-114	4.007	75	2.937	Leu-5	3.989	64	3.739
42	4.357	74	3.068	11-39	3.912	60	3.476	Leu-1	3.932	66	3.086
26	4.143	83	3.588	11-50	3.689	55	3.616	Leu-53	3.807	67	4.113
39	4.111	71	3.175	10-9	3.555	73	2.783	Leu-3	3.761	78	4.303
72	3.970	81	4.623	10-21	3.526	58	3.133	Leu-41	3.689	84	2.975
25	3.892	71	4.001	11-41	3.401	69	3.612	Leu-28	3.497	66	3.587
63	3.434	71	3.533	10-20	3.367	75	3.240	Leu-31	3.296	83	3.525
7	3.434	70	2.692	10-23	3.091	62	2.636	Leu-24	2.833	75	2.427
67	3.434	90	3.168	10-27	3.045	75	3.240				
1	3.401	79	4.234	11-104	2.996	64	3.653				
40	3.091	80	3.674	10-22	2.944	68	3.615				
13	2.996	68	3.547	11-48	2.639	55	2.485				
9	2.944	68	2.785	11-49	2.565	63	2.971				
60	2.890	92	3.447	11-58	2.485	58	2.517				
59	2.708	82	2.936	10-26	2.175	53	2.808				
2	2.398	67	2.904	11-69	2.140	53	2.657				
33	2.140	79	3.062	11-103	2.140	63	3.053				
52	2.001	85	3.647	11-73	2.028	66	1.200				
29	1.856	80	3.415	11-55	1.856	50	2.306				
20	1.281	81	1.656	11-45	1.856	58	2.517				
49	1.253	82	1.967	11-51	1.668	56	1.897				
46	0.3365	80	2.463								
				Site 2[e]	4.605	91	5.160				
				Site 4[e]	1.030	66	0.7098				

[a] Each group of four columns represents data from a different Selex experiment.

[b] The natural logarithm of experimentally determined relative binding constants. DNA from each clone was cut with *Hin*dIII and *Xba*I or *Eco*RI and *Bam*HI, yielding a 35-mer containing the Lrp binding site and the remainder of the vector. The products were end labeled with ^{32}P, incubated for 20 min with Lrp, and fractionated by electrophoresis. The data resemble those shown in Fig. 2. Radioactivity in the free 35-mer and in the 35-mer–Lrp complex was determined with a Betascope blot analyzer. *ilvIH* site 2 DNA was included in each experiment, and the values shown are relative to that for site 2. Relative binding strengths for clones 37, 40, 42, 49, 52, 55, 59, 60, 63, 69, 10-3, 11-39, 11-51, and 11-55 and sites 2 and 4 were measured in experiments using a range of Lrp concentrations. For the remainder of the clones, a single Lrp concentration of 10 nM was used, and the binding strength relative to that of site 2 was calculated from the equation $K_{site\ 2}/K = 100[(100/X_2) - 1]/[(100/X_1) - 1]$, where $K_{site\ 2}$ and K are equilibrium dissociation constants and X_2 and X_1 are the percentages of the total DNA that exist as complex for site 2 DNA and the comparison DNA, respectively. Values are the average of three experiments; standard deviations were less than 15% of the mean for 80% of the isolates and less than 35% for the remainder.

[c] Similarity scores calculated by the procedure of Goodrich et al. (17). Similarity score = 100 × [(score for sequence of interest − baseline score)/(maximum score − baseline score)]. Scores were determined by adding the relevant values at each position of the consensus shown in Table 1. For example, the value for position 1 for a 15-base sequence beginning with C is 60. The baseline score, used to correct for random occurrence of bases in a sequence, is 25% of the sum of all scores. Baseline and maximum scores are 375 and 1,067, respectively.

[d] The natural logarithms of binding energies predicted from the multiple regression analysis of Stormo et al. (36) and Barrick et al. (3), using the matrix data shown in Fig. 7A. Example for clone 55: the sum of the matrix values for the sequence of clone 55 (Fig. 3; sequence CAGGATATTATTCTG) is 1.610. A constant value of 2.94 is added, giving a predicted natural logarithm of binding energy of 4.595. Note that for Leu-5, the sequence used in the alignment was TAGCAGTTTATTCTC rather than the sequence shown in Fig. 5.

[e] The sequences of sites 2 and 4 matching the consensus are TAGAATTTTATTCTG and GAGGATTTTATCGTT, respectively (41).

dromic consensus sequence. It may be that half palindromic sites bind Lrp more weakly than do full palindromic sites and that the stronger full palindromic sites were selected for in these in vitro experiments.

How useful is the matrix in Fig. 7A for predicting Lrp binding sites within DNA of any given sequence? The 63 Lrp binding sites summarized in Table 4 vary in binding strength over a reasonably wide range (about 285-fold). The average natural logarithm of binding energy for this set calculated from the matrix is 3.34 (standard deviation, 0.79; maximum, 4.96; minimum, 1.2). This mean can be compared with a mean of -8.29 ± 3.83 for 30 random sequences (maximum, 0.93; minimum, −14.1). The two means differ by 2.7 standard deviations. It should be noted that the mean for the random sequences is to some degree arbitrary because some of the entries in the matrix were assigned arbitrary values (those marked with asterisks in Fig. 7A) because those bases do not appear at that position in any of the sequences analyzed. The values that we chose for those cases (−2.5, except for position 8, which we assigned −5) were conservative, given the absence of those particular bases among the 63 sequences compared. Next we applied the matrix analysis to naturally occurring sequences, some within coding regions of genes (not expected to have Lrp binding sites) and some known to bind Lrp in vitro. A computer program was written that calculates natural logarithms of binding energies of successive 15-bp stretches of DNA by using the matrix in Fig. 7A and that displays scores higher than an assigned cutoff value. Predicted binding energies having a natural logarithm greater than 4 (note that a strong binding site, such as *ilvIH* site 2, has a measured natural

A

Position	1	2	3	4	5	6	7	8
G	-1.36	-1.25	0.0	-0.96	0.12	-0.36	-0.30	-5*
A	-1.44	0.0	-1.26	-0.00	0.91	-0.09	0.53	-5*
T	0.0	-2.5*	-2.5*	0.0	0.0	0.0	0.0	0.0
C	-0.14	-0.34	-2.5*	0.15	0.35	-0.35	-2.13	-5*
Consensus	Y	A	G	H	A	W	A	T

Position	15	14	13	12	11	10	9
G	-0.14	-0.34	-2.5*	0.15	0.35	-0.35	-2.13
A	0.0	-2.5*	-2.5*	0.0	0.0	0.0	0.0
T	-1.44	0.0	-1.26	-0.00	0.91	-0.09	0.53
C	-1.36	-1.25	0.0	-0.96	0.12	-0.36	-0.30
Consensus	R	T	C	D	T	W	T

B

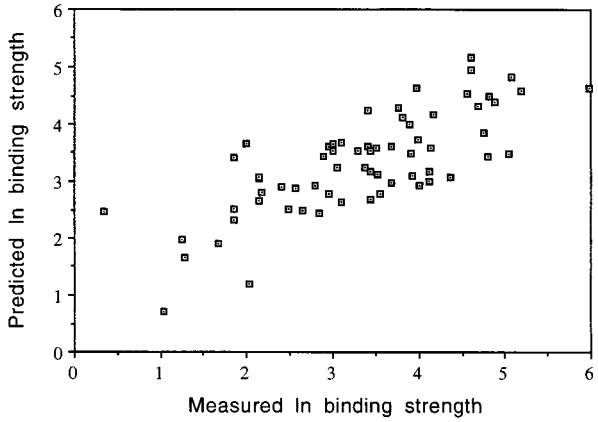


FIG. 7. (A) Binding energies (natural logarithms) associated with each position of the consensus sequence, estimated by using the procedures of Stormo et al. (36) and Barrick et al. (3). To estimate ln binding energy for any particular 15-base sequence, sum the relevant values and add 2.94. Values highlighted by asterisks were arbitrarily assigned for reasons that follow. Position 8 in the consensus is always an AT base pair; by assigning −5 to all positions except T, only sequences having a T at position 8 are considered significant. For the other positions marked by asterisks, none of 63 sequences analyzed had the indicated nucleotide at that position, so that the program assigned neither a positive nor a negative value. We assigned an arbitrary value of −2.5 to reflect the obvious bias in the Selex experiments against these nucleotides. These arbitrary assignments were made to make the matrix more useful in predicting Lrp binding sites in chromosomal DNA. Searching both strands of a sequence will ensure that all sequences will be fairly considered that have an AT base pair at position 8. Y = C or T; H = not G; W = A or T; D = not C; R = A or G. (B) Comparison of experimentally determined relative binding energies with those predicted from the methods of Stormo et al. (36) and Barrick et al. (3) (natural logarithms in both cases).

logarithm of binding energy of 4.6) are rare: none were found within the *E. coli trpEDCBA* or *araBAD* operon (accession numbers V00372 and M15263, respectively), and only 9 were found within a 100,000-bp region at min 76 of the *E. coli* chromosome (accession number U00039). Thus, it seems significant that two of the *E. coli* operons controlled by Lrp, *ilvIH* site 2 and *pap*, have predicted natural logarithms of binding energies greater than 4 and that the relevant sites are protected by Lrp in in vitro footprinting experiments (26, 40). Making predictions based on calculated scores less than 4 does not seem warranted at this time. Calculated scores between 3 and 4 are found within DNAs showing Lrp-induced footprints near *fim* (14), *lysU* (16), and *lrp* (43) (accession numbers M11774, M30630, and U02275, respectively), but they are also found within structural genes of the *trp* and *ara* operons (two and four occurrences, respectively). A complicating factor in this analysis is the fact that Lrp is known in some cases to bind cooperatively to multiple sites, for example, to sites upstream of the

TABLE 5. Comparison of consensus sequences

Sequence[a]																Reference
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
Y	A	G	H	A	W	A	T	T	W	T	D	C	T	R		This study
	A	g	a	A	T	T	T	T	A	T	t	c	T			42
						T	T	T	A	T	T	C	t	N	a A T	32

[a] Y = C or T, H = not G, W = A or T, D = not C, and R = A or G.

ilvIH operon of *E. coli* (40). Thus, strong binding may result from cooperative interaction of Lrp with sites that individually are relatively weak. In the case of the *ilvIH* operon of *E. coli*, the cooperative interactions occur between sites that are spaced not more than about 30 bp apart. We searched by computer for scores above zero that were separated by 50 bp or less and looked for patterns of relatively weak closely spaced binding sites. Clusters of closely spaced relatively weak predicted sites correlate with known binding sites 3, 4, 5, and 6 for the *E. coli ilvIH* operon, but again, similar clusters are found within structural genes of *trp* and *ara*. For the future, better prediction will likely require an improved correlation between predicted and observed binding energies for Lrp binding to 15-mer sequences and more information about the requirements for cooperative binding of Lrp to multiple sites.

Lrp is known to be a homodimer in solution (45). The palindromic nature of the consensus sequence derived here suggests that Lrp binds to DNA as a homodimer in such a way that each monomer binds to a half-site. The stoichiometry of Lrp binding to DNA has not yet been reported, but our preliminary results suggest that a single site binds a dimer (23a).

A most interesting feature of the Lrp regulon is the variety of regulatory mechanisms that result from the interaction of Lrp and leucine. Lrp can act as either an activator or a repressor, and for each of these situations, examples are known in which leucine overcomes the effect of Lrp, leucine is required for the effect of Lrp, or leucine has no effect upon Lrp action (25). The molecular mechanisms underlying these regulatory mechanisms are for the most part not understood. One hypothesis that can explain how leucine might overcome an effect of Lrp in some systems and be required for an effect in others is that Lrp binds to different sequences in the presence and absence of leucine. The results presented here suggest that this is not the case: sequences selected for their ability to bind Lrp were the same or very similar irrespective of whether the selection was performed in the presence or absence of leucine. Equally important, leucine did not stimulate binding of Lrp to any of the sequences that were selected in the presence of leucine. In fact, for each single Lrp binding site that we studied, including individual sites within *ilvIH* DNA, leucine reduced the extent of Lrp binding. The in vitro stimulatory effects of leucine that we and others have observed (10) are most readily explained in terms of effects of leucine upon cooperative binding of Lrp to multiple sites.

ACKNOWLEDGMENTS

This work was supported by grants GM39496 and HG00249 from the National Institutes of Health (to J.M.C. and G.D.S., respectively) and by grant ER61606 to G.D.S. from the Department of Energy. Y. Cui was partly supported by a fellowship from the K.-C. Wong Education Foundation, Hongkong.

We thank Debra A. Willins for purified Lrp, Sarah Calvo for computer programming, John Heumann for help with alignments, and Jeffrey Roberts, P. Hinkle, Rowena Matthews, and Richard Ebright for helpful discussions.

REFERENCES

1. **Austin, E. A., J. C. Andrews, and S. A. Short.** 1989. Selection, characterization, and cloning of *oppI*, a regulator of the *E. coli* oligopeptide permease operon, p. 153. *In* Abstracts of the Papers Presented at the 1989 Meeting on Molecular Genetics of Bacteria and Phages: Prokaryotic Genetic Regulation. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
2. **Ausubel, F. M., R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith, and K. Struhl (ed.).** 1989. Current protocols in molecular biology, p. 3.1.1–3.4.10. Greene Publishing Associates, New York.
3. **Barrick, D., K. Villanueba, J. Childs, R. Kalil, T. D. Schneider, C. E. Lawrence, L. Gold, and G. D. Stormo.** 1994. Quantitative analysis of ribosome binding sites in *E. coli*. Nucleic Acids Res. **22:**1287–1295.
4. **Berg, O. G.** 1988. Selection of DNA binding sites by regulatory proteins; the LexA protein and the arginine repressor use different strategies for functional specificity. Nucleic Acids Res. **16:**5089–5105.
5. **Berg, O. G., and P. A. von Hippel.** 1988. Selection of DNA binding sites by regulatory proteins. II. The binding specificity of cyclic AMP receptor protein. J. Mol. Biol. **200:**709–723.
6. **Blackwell, T. K., L. Kretzner, E. M. Blackwood, R. N. Eisenman, and H. Weintraub.** 1990. Sequence-specific DNA binding by the c-Myc protein. Science **250:**1149–1151.
7. **Blackwell, T. K., and H. Weintraub.** 1990. Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. Science **250:**1104–1110.
8. **Blomfield, I. C., P. J. Calie, K. J. Eberhardt, M. S. McClain, and B. I. Eisenstein.** 1993. Lrp stimulates phase variation of type 1 fimbriation in *Escherichia coli* K-12. J. Bacteriol. **175:**27–36.
9. **Braaten, B. A., J. V. Platko, M. van Woude, F. K. de Graaf, J. M. Calvo, and D. A. Low.** 1992. Leucine responsive regulatory protein (Lrp) controls the expression of both the Pap and Fan pili operons in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **89:**4250–4254.
10. **Calvo, J. M., and R. G. Matthews.** 1994. The leucine-responsive regulatory protein (Lrp), a global regulator of metabolism in *Escherichia coli*. Microbiol. Rev. **58:**466–490.
11. **Chittenden, T., D. M. Livingston, and W. G. Kaelin.** 1991. The T/E1A-binding domain of the retinoblastoma product can interact selectively with a sequence-specific DNA-binding protein. Cell **65:**1073–1082.
12. **Devereux, J., P. Haeberli, and O. Smithies.** 1984. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. **12:**387–395.
13. **Ernsting, B. R., M. R. Atkinson, A. J. Ninfa, and R. G. Matthews.** 1992. Characterization of the regulon controlled by the leucine-responsive regulatory protein in *Escherichia coli*. J. Bacteriol. **174:**1109–1118.
14. **Gally, D. L., T. J. Rucker, and I. C. Blomfield.** 1994. The leucine-responsive regulatory protein binds to the *fim* switch to control phase variation of type 1 fimbrial expression in *Escherichia coli* K-12. J. Bacteriol. **176:**5665–5672.
15. **Gazeau, M., F. Delort, P. Dessen, S. Blanquet, and P. Plateau.** 1992. *Escherichia coli* leucine-responsive regulatory protein Lrp controls lysyl-tRNA synthetase expression. FEBS Lett. **300:**254–258.
16. **Gazeau, M., F. Delort, M. Fromant, P. Dessen, S. Blanquet, and P. Plateau.** 1994. Structure-function relationship of the Lrp-binding region upstream of *lysU* in *Escherichia coli*. J. Mol. Biol. **241:**378–389.
17. **Goodrich, J. A., M. L. Schwartz, and W. R. McClure.** 1990. Searching for and predicting the activity of sites for DNA binding proteins: compilation and analysis of the binding sites for *Escherichia coli* integration host factor (IHF). Nucleic Acids Res. **17:**4993–5000.
18. **Haney, S. A., J. V. Platko, D. L. Oxender, and J. M. Calvo.** 1992. Lrp, a leucine-responsive protein, regulates branched-chain amino acid transport genes in *Escherichia coli*. J. Bacteriol. **174:**108–115.
19. **Heumann, J. M., A. S. Lapedes, and G. D. Stormo.** 1994. Neural networks for determining protein specificity and multiple alignment of binding sites, p. 188–194. *In* R. Altman (ed.), Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park, Calif.
20. **Lin, R., R. D'Ari, and E. B. Newman.** 1992. p*lac*Mu insertions in genes of the leucine regulon: extension of the regulon to genes not regulated by leucine. J. Bacteriol. **174:**1948–1955.
21. **Lin, R., B. Ernsting, I. N. Hirshfield, R. G. Matthews, F. C. Neidhardt, R. L. Clark, and E. B. Newman.** 1992. The *lrp* gene product regulates expression of *lysU* in *Escherichia coli* K-12. J. Bacteriol. **174:**2779–2784.
22. **Lin, R. T., R. D'Ari, and E. B. Newman.** 1990. The leucine regulon of *E. coli* K-12: a mutation in *rblA* alters expression of L-leucine-dependent metabolic operons. J. Bacteriol. **172:**4529–4535.
23. **Mavrothalassitis, G., G. Beal, and T. S. Papas.** 1990. Defining target sequence(s) of a DNA binding protein by random selection and PCR: determination of the GCN4 binding sequence repertoire. DNA Cell Biol. **9:**783–788.

23a. **Midkiff, M., Q. Wang, and J. M. Calvo.** Unpublished data.

24. **Mossing, M. C., and M. T. Record.** 1985. Thermodynamic origins of specificity in the *lac* repressor-operator interaction. Adaptability in the recognition of mutant operator sites. J. Mol. Biol. **186:**295–305.
25. **Newman, E. B., R. D'Ari, and R. T. Lin.** 1992. The leucine-Lrp regulon in *E. coli*: a global response in search of a raison d'etre. Cell **68:**617–619.
26. **Nou, X., B. Skinner, B. Braaten, L. Blyn, D. Hirsch, and D. Low.** 1993. Regulation of pyelonephritis-associated pili phase-variation in *Escherichia coli*: binding of the PapI and the Lrp regulatory proteins is controlled by DNA methylation. Mol. Microbiol. **7:**545–553.
27. **O'Neill, M. C.** 1989. Consensus methods for finding and ranking DNA binding sites. Application to *Escherichia coli* promoters. J. Mol. Biol. **207:** 301–310.
28. **Perkins, A. S., R. Fishel, N. A. Jenkins, and N. G. Copeland.** 1991. *Evi-1*, a murine zinc finger proto-oncogene, encodes a sequence-specific DNA-binding protein. Mol. Cell Biol. **11:**2665–2674.
29. **Platko, J. V., and J. M. Calvo.** 1993. Mutations affecting the ability of *Escherichia coli* Lrp to bind DNA, activate transcription, or respond to leucine. J. Bacteriol. **175:**1110–1117.
30. **Platko, J. V., D. A. Willins, and J. M. Calvo.** 1990. The *ilvIH* operon of *Escherichia coli* is positively regulated. J. Bacteriol. **172:**4563–4570.
31. **Pollock, R., and R. Treisman.** 1990. A sensitive method for the determination of protein-DNA binding specificities. Nucleic Acids Res. **18:**6197–6204.
32. **Rex, J. H., B. D. Aronson, and R. L. Somerville.** 1991. The *tdh* and *serA* operons of *E. coli*: mutational analysis of the regulatory elements of leucine-responsive genes. J. Bacteriol. **173:**5944–5953.
33. **Ricca, E., D. A. Aker, and J. M. Calvo.** 1989. A protein that binds to the regulatory region of the *ilvIH* operon of *Escherichia coli*. J. Bacteriol. **171:** 1658–1664.
34. **Sambrook, J., E. F. Fritsch, and T. Maniatis.** 1989. Molecular cloning: a laboratory manual, 2nd ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
35. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:**5463–5467.
36. **Stormo, G. D., T. D. Schneider, and L. Gold.** 1986. Quantitative analysis of the relationship between nucleotide sequence and functional activity. Nucleic Acids Res. **14:**6661–6679.
37. **Takeda, Y., A. Sarai, and V. M. Rivera.** 1989. Analysis of the sequence-specific interactions between Cro repressor and operator DNA by systematic base substitution experiments. Proc. Natl. Acad. Sci. USA **86:**439–443.
38. **Thiesen, H.-J., and C. Bach.** 1990. Target detection assay (TDA): a versatile procedure to determine DNA binding sites as demonstrated on SP1 protein. Nucleic Acids Res. **18:**3203–3209.
39. **Tuerk, C., and L. Gold.** 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science **249:**505–510.
40. **Wang, Q., and J. M. Calvo.** 1993. Lrp, a global regulatory protein of *E. coli*, binds cooperatively to multiple sites and activates transcription of *ilvIH*. J. Mol. Biol. **229:**306–318.
41. **Wang, Q., and J. M. Calvo.** 1993. Lrp, a major regulatory protein in *Escherichia coli*, bends DNA and can organize the assembly of a higher-order nucleoprotein structure. EMBO J. **12:**2495–2501.

41a. **Wang, Q., and J. M. Calvo.** Unpublished data.

42. **Wang, Q., M. Sacco, E. Ricca, C. T. Lago, M. DeFelice, and J. M. Calvo.** 1993. Organization of Lrp binding sites upstream of *ilvIH* in *Salmonella typhimurium*. Mol. Microbiol. **7:**883–891.
43. **Wang, Q., J. Wu, D. Friedberg, J. Platko, and J. M. Calvo.** 1994. Regulation of the *Escherichia coli lrp* gene. J. Bacteriol. **176:**1831–1839.
44. **Willins, D. A., and J. M. Calvo.** 1992. In vitro transcription from the *Escherichia coli ilvIH* promoter. J. Bacteriol. **174:**7648–7655.
45. **Willins, D. A., C. W. Ryan, J. V. Platko, and J. M. Calvo.** 1991. Characterization of Lrp, an *Escherichia coli* regulatory protein that mediates a global response to leucine. J. Biol. Chem. **266:**10768–10774.