One Exon of the Human LSF Gene Includes Conserved Regions Involved in Novel DNA-Binding and Dimerization Motifs

MARGARET K. SHIRRA,^{1,2} QUAN ZHU,^{1,2} HUI-CHUAN HUANG,^{1,2†}
DAVID PALLAS,^{3,4} AND ULLA HANSEN^{1,2*}

Division of Molecular Genetics¹ and Department of Cellular and Molecular Biology,³ Dana-Farber Cancer Institute, and Department of Microbiology and Molecular Genetics² and Department of Pathology,⁴ Harvard Medical School, Boston, Massachusetts 02115

Received 7 September 1993/Returned for modification 8 November 1993/Accepted 5 May 1994

The transcription factor LSF, identified as a HeLa protein that binds the simian virus 40 late promoter, recognizes direct repeats with a center-to-center spacing of 10 bp. The characterization of two human cDNAs, representing alternatively spliced mRNAs, provides insight into the unusual DNA-binding and oligomerization properties of LSF. The sequence of the full-length LSF is identical to that of the transcription factors α CP2 and LBP-1c and has similarity to the *Drosophila* transcription factor Elf-1/NTF-1. Using an epitope-counting method, we show that LSF binds DNA as a homodimer. LSF-ID, which is identical to LBP-1d, contains an in-frame internal deletion of 51 amino acids resulting from alternative mRNA splicing. Unlike LSF, LSF-ID did not bind LSF DNA-binding sites. Furthermore, LSF-ID did not affect the binding of LSF to DNA, suggesting that the two proteins do not interact. Of three short regions with a high degree of homology between LSF and Elf-1/NTF-1, LSF-ID lacks two, which are predicted to form β -strands. Double amino acid substitutions in each of these regions eliminated specific DNA-binding activity, similarly to the LSF-ID deletion. The dimerization potential of these mutants was measured both by the ability to inhibit the binding of LSF to DNA and by direct protein-protein interaction studies. Mutations in one homology region, but not the other, functionally eliminated dimerization.

Regulation of the initiation of transcription in mammalian cells relies substantially on the activities of sequence-specific DNA-binding proteins. The functional analysis of cDNAs for transcription factors has therefore provided a means for understanding molecular mechanisms by which transcriptional activation and repression occur, both in vitro and in vivo. Such cDNA clones allow dissection of oligomerization, DNA-binding, and transcriptional activation activities.

Transcription factors often bind DNA as oligomers, allowing recognition of repeated sequences in the DNA. Most such DNA-binding sites consist of inverted repeated sequences, which are readily recognized by symmetric dimeric or tetrameric proteins. On the other hand, proteins have developed a variety of strategies to bind direct repeats. The eukaryotic heat shock transcription factors are trimers, which allow recognition of both inverted and direct repeats with defined spacings (29, 35). The phage transcriptional regulator λ cII forms tetramers to interact with direct repeats, although the exact mechanism of interaction is not clear (17). Finally, dimeric proteins that bind direct repeats are relatively uncommon. As indicated above, when the dimerization interface involves the same region in each monomer, creating a twofold rotational axis of symmetry, the two DNA-binding domains in the dimer are positioned to recognize inverted repeats on DNA. In contrast, dimeric proteins that recognize direct repeats either must structurally separate such a twofold symmetric dimerization domain from the DNA-binding domains

or must dimerize in a head-to-tail fashion. In the latter case, two different surfaces of the protein may be involved in dimeric interactions. In such instances, a mechanism to prevent reiterated head-to-tail oligomerization by the dimerization region is required. The first mechanism of binding directly repeated DNA sequences by a protein dimer is exemplified by COUP-TF (8), vitamin D₃ receptor (12), and the yeast α 2 repressor (30). All seem to contain flexible linkers between the dimerization and DNA-binding domains, as the spacing between the half-sites can vary and each can bind either direct or inverted repeats. However, interaction of the α 2 repressor with the MCM1 protein in vivo restricts the flexibility of the repressor, forcing it to recognize only inverted repeats, as occur in the natural promoter (34). Examples of head-to-tail dimers have not yet been described. As we discuss here, LSF may represent this type of binding.

LSF, a mammalian transcription factor, was initially described as a protein from HeLa cells that stimulated transcription from the major late simian virus 40 promoter in vitro (21). LSF specifically binds two sites in the simian virus 40 promoter region: LSF-280, a bipartite site centered 44 bp upstream of the major late initiation site, and LSF-GC, consisting of direct repeats of the sequence GGGCGG separated by 4 bp. The centers of the repeats in LSF-GC are separated by 10 bp, suggesting that oligomeric LSF binds to the site on a single face of the DNA double helix.

To study further the DNA-binding and oligomeric properties of LSF, we obtained two cDNAs for this transcription factor. These two cDNAs result from alternative splicing and encode LSF and LSF-ID. We show that LSF binds to LSF-280 DNA as a dimeric complex. In contrast, LSF-ID, which has an internal deletion of 51 amino acids from the central portion of the protein, does not bind LSF DNA-binding sites and, in

* Corresponding author. Mailing address: Division of Molecular Genetics, Dana-Farber Cancer Institute, B411, 44 Binney St., Boston, MA 02115. Phone: (617) 375-8211. Fax: (617) 375-8221.

† Present address: Department of Cellular and Molecular Physiology, Harvard Medical School, Boston, MA 02115.

addition, does not affect the binding in vitro of LSF to its DNA sites. The phenotypes of two double-substitution mutants of LSF, altered within this region of the protein, suggest that dimerization and possibly DNA-binding functions are localized, in part, within this small segment of LSF.

MATERIALS AND METHODS

Cloning of LSF cDNAs. LSF was purified from a whole-cell extract (25) of HeLa cells by fractionation through a heparin column and several successive DNA-affinity columns containing two different LSF DNA-binding sites, as previously described (19). The final step of purification involved preparative two-dimensional electrophoresis, separating LSF by isoelectric point in the first dimension and by sodium dodecyl sulfate (SDS)-polyacrylamide gel electrophoresis (10% acrylamide) in the second dimension (27). Gradients of approximately pH 4 to 8 were obtained by mixing 400 μ l of pH 3.5 to 10 ampholine (LKB) with 150 μ l of pH 6 to 8 ampholine (LKB) per 10 ml of gel solution. LSF migrated as several isoelectric variants, consistent with possible posttranslational modification. The proteins in the two-dimensional gels were electrophoretically transferred to nitrocellulose membranes at 0.5 A for 3.5 h in an electroblotting apparatus (Hoefer), visualized by staining with 1% Ponceau S (1), and then excised. Except for high-pressure liquid chromatography (HPLC), internal amino acid sequence analysis was done by the method of Abersold et al. (1). Resultant tryptic digestion mixtures were fractionated and eluted with a gradient essentially as described previously (37) on a Hewlett-Packard 1090 HPLC equipped with a 1040 diode array detector and a Vydac C₁₈ column. Separated peptides were applied to an Applied Biosystems model 477A pulsed-liquid protein sequencer equipped with an on-line model 120A HPLC using the manufacturer's fast cycle program. Six tryptic peptide sequences were obtained (see Fig. 1).

Two pairs of degenerate oligonucleotides, reflecting adjacent amino acid sequences in one of the peptides, were synthesized (I = inosine): 1, AA (T/C) CA (A/G) GGICA (A/G) TCITA (T/C) GA (A/G) AT (NQGQSYEI-sense); 2, AA (T/C) CA (A/G) GGICA (A/G) AG (T/C) TA (T/C) GA (A/G) AT (NQGQSYEI-sense); 3, TAIGTIAGIGT (T/C) TC (A/G) TC (A/G) TG (HDETLY-antisense); and 4, TAIGT (T/C) AA IGT (T/C) TC (A/G) TC (A/G) TG (HDETLY-antisense). Duplicate filters from a π H3M-based plasmid library prepared with cDNA from the human T-cell tumor line HPB-ALL (a gift from Brian Seed) (3, 31) were screened separately with pairs of oligonucleotides, either 1 and 2 or 3 and 4. Filters were prehybridized in 6 $\times$ SSC (0.9 M NaCl plus 0.09 M sodium citrate, pH 7.0)–5 $\times$ Denhardt's solution–1% SDS–0.5% sodium PP_i–100 μ g of tRNA per ml at 37°C overnight. Phosphorylated oligonucleotides, at 125 pg/ml per degenerate oligonucleotide, were added and allowed to hybridize at 37°C for 2 days. All filters were washed twice for 15 min at room temperature with 6 $\times$ SSC–0.05% sodium PP_i. Filters probed with oligonucleotide pair 1 and 2 were washed twice at 47°C for 15 min with 3 M tetramethylammonium chloride–50 mM Tris-HCl (pH 8.0)–2 mM EDTA. Filters probed with oligonucleotide pair 3 and 4 were washed twice at 43°C in the same solution. From 6 $\times$ 10⁵ colonies, two clones positive for both pairs of oligonucleotides were selected and purified (5). Chain-termination sequencing (Sequenase; U.S. Biochemical Corp.) was used to sequence at least one strand from both clones, using appropriate primers. Discrepancies and additional sequences between our clones and the published human α CP2 sequence (23) were verified by sequencing both strands. All

sequencing reactions were performed with the original π H3M constructs.

Plasmid construction. The expression clone pET-LSF contains sequences from the LSF cDNA and pET-11c (Novagen), which contains, for our purposes, a T7 promoter and transcription termination sequences. The *Bam*HI site of pET-11c was converted to an *Xho*I site with translation termination codons in all three frames by insertion of

GATCTCGAGGCGGTACCTAGCTAGCTAGT
AGCTCCGCCATGGATCGATCGATCACTAG.

pET-11q contained the insert oriented as shown above between the T7 promoter and T7 termination sequences, and pET-11qr contained the insert in the reverse orientation. pET-LSF-ID Δ 24 was constructed by inserting the *Nhe*I-*Xho*I fragment of π H3M-LSF-ID into the *Nhe*I-*Xho*I backbone of pET-11qr. pET-LSF Δ 24 was constructed by exchanging the *Nhe*I-*Nsi*I fragment of π H3M-LSF with the *Nhe*I-*Nsi*I fragment of pET-LSF-ID Δ 24. pET-LSFqr was constructed by removing the LSF sequences between *Nhe*I and *Nsi*I of pET-LSF-ID Δ 24 and replacing them with LSF sequences from the *Bgl*I-*Nsi*I fragment of π H3M-LSF and the linker

CTAGCCGCGGCGGATCCGGGA
GGCGCCGCCTAGGC.

pET-LSF-ID was similarly constructed by ligating the *Nhe*I-*Nsi*I backbone of pET-LSF-ID Δ 24 to the *Bgl*I-*Nsi*I fragment of π H3M-LSF-ID with the same oligonucleotides. For pET-LSF, the *Sph*I-*Xho*I fragment containing LSF from pET-LSFqr was exchanged with the *Sph*I-*Xho*I backbone from pET-11q. To generate pGEX-2TK-LSF and pGEX-2TK-LSF-ID, the DNA sequences encoding LSF and LSF-ID were excised from pET-LSF and pET-LSF-ID, respectively, by digestion with *Bam*HI and *Xho*I. These fragments were ligated, with the coding region in frame, to pGEX-2TK (20) that had been digested with *Bam*HI and *Eco*RI, after the *Xho*I site of LSF and the *Eco*RI site of the vector were filled in with the Klenow fragment of *Escherichia coli* DNA polymerase (5).

Reverse transcriptase-PCR. Total RNA from HeLa and HPB-ALL cells was prepared by using RNazol (Tel-Test, Inc., Friendswood, Tex.) according to the manufacturer's instructions. RNA was then treated with RNase-free DNase (Worthington, Freehold, N.J.) (5). A 30- μ g amount of total RNA and 500 ng of 3' primer (see below) were annealed at 70°C and then extended at 37°C for 90 min by the addition of 200 U of SuperScript RNase H-reverse transcriptase (GIBCO-BRL) in 50 mM Tris-HCl (pH 8.3)–75 mM KCl–3 mM MgCl₂–10 mM dithiothreitol (DTT)–500 μ M deoxynucleoside triphosphates (dNTPs)–40 U of RNasin (Promega). Control reaction mixtures lacked reverse transcriptase. One-fifth of the cDNA was amplified in a 100- μ l PCR mixture containing 20 pmol of 5' and 3' primers (see below), 400 μ M dNTPs, 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 1.5 mM MgCl₂, 0.01% (wt/vol) gelatin, and 1.25 U of *Taq* polymerase with 30 cycles of the following protocol: 1 min at 94°C, 1 min at 50°C, and 1 min at 72°C. All the PCR components were supplied by Perkin-Elmer/Cetus. Amplified DNAs were electrophoresed through a 2% agarose gel containing 0.05 μ g of ethidium bromide per ml (5). Coamplification of plasmid DNA was performed by using 1 ng of pET-LSF and 1 ng of pET-LSF-ID as described above. The following primers were used: 5' primer, GAGCATCAGC AGCTAGAGGGG, nucleotides (nt) 1088 to 1108 (sense strand, Fig. 1A), and 3' primer, CAGTGATTCTGACA AAC, nt 1888 to 1871 (antisense strand, Fig. 1A). The relative abundance of the mRNAs was determined by scanning a photograph of the stained gel (UMAX UC630 scanner) into

Adobe Photoshop and using ImageQuant (Molecular Dynamics) software to quantitate the bands.

PCR with genomic DNA. A 10- μ g sample of genomic HeLa DNA (5) was amplified under the reaction conditions described above, and 380 μ g of single-strand-binding protein (Pharmacia) per ml was added, using the following protocol: 2 cycles of 2 min at 94°C, 1 min at 50°C, and 2 min at 72°C followed by 28 cycles of 1 min at 94°C, 1 min at 50°C, and 2 min at 72°C. The following primers were used (see Fig. 1A): 5' outer, CAAAGAGGACATCTGTGTTTATTCA (nt 1230 to 1254); 5' inner, CGGATCCGTGCACTGTATTAGC (nt 1256 to 1270 plus *Bam*HI site); 3' outer, TTTTGTCTTCTGTCTGCACCTTT (nt 1435 to 1412); and 3' inner, CGGATTCTTGAAACTTTGATCTGGC (nt 1408 to 1389 plus *Eco*RI site). A 1-ng sample of pET-LSF was also amplified in the absence of single-strand-binding protein by using the same protocol and oligonucleotides. A 40- μ l sample from each PCR mixture was precipitated and digested with 22 U of *Pvu*II (New England Biolabs) for 3 h. Amplified DNAs were electrophoresed through a 1% agarose gel containing 0.05 μ g of ethidium bromide per ml (5).

In vitro transcription-translation reactions. Proteins were synthesized in reticulocyte lysate or wheat germ extract by using the coupled transcription-translation systems (TNT) from Promega. A 500-ng amount of total DNA was added to 25- μ l reaction volumes, and reactions were performed according to manufacturer's instructions, using [³⁵S]methionine.

GST fusion proteins. Glutathione S-transferase (GST)-LSF fusion proteins were expressed and purified essentially as described previously (20), except that $N_{100}E_{50}TN_{0.5}^*$ (20 mM Tris [pH 8.0], 100 mM NaCl, 50 mM EDTA, 0.5% Nonidet P-40 [NP-40], 100 μ g of leupeptin per ml, 100 μ g of aprotinin per ml, 1 mM phenylmethylsulfonyl fluoride) was substituted for NETN. GST proteins were eluted from glutathione-Sepharose beads (Pharmacia) with freshly prepared 20 mM glutathione in 100 mM Tris (pH 8.0)–120 mM NaCl.

Band mobility shift assays. GST proteins were added to 15- μ l reaction mixtures containing 15 mM Tris-HCl (pH 8.0), 20 mM HEPES (*N*-2-hydroxyethylpiperazine-*N'*-2-ethanesulfonic acid) (pH 7.9), 20% glycerol, 2% polyvinyl alcohol, 0.2 mM EDTA, 135 mM KCl-NaCl, 1 mM DTT, and 67 μ g of poly[d(I-C) · d(I-C)] per ml with 500 pg of radiolabeled DNA containing the LSF-280 binding site (probe) (19), and the mixtures were incubated for 30 min at room temperature prior to gel analysis. A 1- μ l amount of protein synthesized with the TNT system was added to 14 μ l containing 250 pg of probe in 8 mM Tris-HCl (pH 8.0)–20 mM HEPES (pH 7.9)–20% glycerol–2% polyvinyl alcohol–0.2 mM EDTA–140 mM KCl–1 mM DTT–3.3 μ g of poly[d(I-C) · d(I-C)] per ml, and the mixture was incubated for 15 min at room temperature prior to gel analysis. A 1- μ l amount of affinity-purified rabbit α -pepLSF1-1 antibody (41) at 800 ng/ μ l in phosphate-buffered saline (137 mM NaCl, 2.7 mM KCl, 4.3 mM Na₂HPO₄ · 7H₂O, 1.4 mM KH₂PO₄) was added as indicated in the legend to Fig. 5. The antiserum was raised against a synthetic peptide containing amino acids 7 to 21 of LSF (41). For analysis of the LSF substitution mutants, 2 μ l of proteins from the TNT system was added to 18 μ l containing 250 pg of probe in 11 mM Tris-HCl (pH 8.0)–11% glycerol–2.2% polyvinyl alcohol–0.11 mM EDTA–110 mM KCl–1.1 mM DTT–5.5 μ g of poly[d(I-C) · d(I-C)] per ml, and the mixture was incubated as described above. Following incubation, the samples were electrophoresed through a 5% polyacrylamide gel (60:1 acrylamide–bisacrylamide) containing 44.5 mM Tris-base, 44.5 mM boric acid, and 1 mM EDTA. Dried gels were visualized with a Molecular Dynamics model 400E PhosphorImager. A sheet of XAR film

(Kodak) was used to block exposure from the ³⁵S-labelled proteins.

Site-directed mutagenesis. Recombinant PCR (16) was used to amplify segments of LSF containing specific mutations. The following primers were used (see Fig. 1A): A, CCCGATG TCTGTGGGTATAATCG (nt 1253 to 1275); B, ACCTTC AAGGAGAATGAAAACGG (nt 1334 to 1356); C, CTCC TTGAAGGTATCT(T/C)TTTGTCTCGGAATGGACCCC (nt 1345 to 1308); D, CACTCGGCCAGCTGCCTGATC(C/G)AAGTTTTCAAGCCC (nt 1376 to 1411); E, GCAGCTG GCCGAGTGTAAAGTGC (nt 1391 to 1369); and F, GGG AACTGTTGAAGCCAGG (nt 1583 to 1565). Underlined sequences represent altered nucleotides. Primary PCRs were performed with primer pairs A and C (reaction 1), A and E (reaction 2), B and F (reaction 3), and D and F (reaction 4) as described above, substituting 200 μ M dNTPs for 400 μ M dNTPs and using the following protocol: 5 cycles of 1.5 min at 94°C, 1 min at 45°C, and 1 min at 72°C followed by 20 cycles of 1 min at 94°C, 1 min at 45°C, and 1 min at 72°C and succeeded by a 5-min extension at 72°C. Reaction products were purified by agarose gel electrophoresis. Secondary reactions were performed with primer pair A and F in the presence of equal molar amounts of products from either reactions 1 and 3 or reactions 2 and 4, using the following protocol: 5 cycles of 1 min at 94°C, 1 min at 32°C, and 2 min at 37°C followed by 15 cycles of 1 min at 94°C, 1 min at 52°C, and 45 s at 72°C, with a final 5-min extension at 72°C. The products of the secondary reactions were digested with *Avr*II and *Nco*I and substituted for the corresponding sequences in pET-LSF. DNA sequence analysis was used to identify mutant plasmids.

GST-LSF oligomerization assay. Bacterial extracts containing GST-LSF were obtained as described above except that NP-40 was omitted from $N_{100}E_{50}TN_{0.5}^*$ and 1 mM EDTA was substituted for 50 mM EDTA. For each reaction, extract containing approximately 1 μ g of GST-LSF was incubated with 30 μ l of glutathione-Sepharose beads (Pharmacia) and approximately 400 μ l of $N_{100}ET$ (100 mM NaCl, 1 mM EDTA, 20 mM Tris-HCl [pH 8.0]) for 30 min at room temperature. The beads were pelleted and washed twice with 1 ml of $N_{100}ET$. The beads bound with GST-LSF were resuspended in 200 μ l of binding buffer containing 10 mM Tris-HCl [pH 7.9], 10% glycerol, 0.1 mM EDTA, 0.5 mM DTT, 100 mM KCl, 0.1% NP-40, and 150 μ g of bovine serum albumin per ml (Sigma). A 2- μ l sample of protein synthesized with the TNT system was added to the beads, and the mixture was incubated for 1 h at room temperature. The beads were washed four times with either $N_{100}ETN_{0.5}$ (0.5% NP-40 added to $N_{100}ET$) or $N_{500}ETN_{0.5}$ (500 mM NaCl substituted for 100 mM NaCl in $N_{100}ETN_{0.5}$). Proteins bound to the beads were analyzed by SDS-polyacrylamide electrophoresis, and radiolabelled proteins were visualized with the PhosphorImager.

Nucleotide sequence accession numbers. The GenBank nucleotide sequence accession number for LSF cDNA is U03494, and the nucleotide sequence accession number for LSF-ID is U03495.

RESULTS

Cloning of LSF cDNAs. To pursue our studies of the DNA-binding and transcriptional activation properties of LSF (see references 19 and 38), cDNAs for LSF were cloned. Degenerate oligonucleotides reflecting one of the six tryptic peptides obtained by microsequencing of LSF purified from HeLa cells were synthesized. The inosine-containing oligonucleotides were used to screen a cDNA plasmid library from the human T-cell tumor line HPB-ALL (3, 31). Two positive

A.

[illegible]

B.

FIG. 1. (A) Nucleic acid and protein sequences of LSF and LSF-ID. Boxed amino acids indicate peptide sequences obtained from microsequencing tryptic peptides derived from purified HeLa LSF. The only discrepancy in sequence was Ser-16, sequenced as a Thr. The nucleotide sequence of the cDNA for LSF-ID is identical to that of the cDNA for LSF with the following exceptions. (i) The 5' end differs as indicated (lowercase letters). Identities between the two noncoding regions (*) are indicated. (ii) LSF-ID cDNA lacks 153 bp, as indicated (double underline). (iii) LSF-ID does not contain amino acid Gln-491 (underlined). (iv) The LSF-ID cDNA ends at bp 2129 of the cDNA encoding LSF. Both of these clones differ from the published sequence for the human α CP2 (23) as follows (base pair number in LSF sequence, LSF sequence $\leftrightarrow$ α CP2 sequence, and deletions Δ) are given, with the amino acid change, if any, shown in parentheses): 481 G $\leftrightarrow$ T; 513 G $\leftrightarrow$ A; 547 G $\leftrightarrow$ A; 651 Δ $\leftrightarrow$ G; 580 Δ $\leftrightarrow$ T; 966 G $\leftrightarrow$ C (Gly-92 $\leftrightarrow$ Ala); 982 T $\leftrightarrow$ C; 1000 G $\leftrightarrow$ A; 1437 C $\leftrightarrow$ T (Thr-249 $\leftrightarrow$ Met); 2162 CAG $\leftrightarrow$ Δ (Gln-491 $\leftrightarrow$ Δ); 2208 GCGT $\leftrightarrow$ ACCG; 2268 G $\leftrightarrow$ A; 2278 Δ $\leftrightarrow$ A; 2299 ACC $\leftrightarrow$ ACG; 2360 TG $\leftrightarrow$ Δ . (B) Schematic comparison of cDNAs encoding LSF and LSF-ID. The nucleotide sequences unique to LSF (diagonals) and the segment of LSF-ID which is identical to a portion of a transposable element (vertical stripes) (see text) are shown. The 5' noncoding regions common to both clones (cross-hatching), the portion of LSF that is similar to Elf-1/NTF-1 (stippling), and the regions most similar to Elf-1/NTF-1 (solid boxes) are indicated. The 10 consecutive glutamine residues (Q) are indicated. The region from bp 2392 through 2418 of the larger clone is an artifact of the cDNA cloning, resulting from self-priming through 6-bp inverted repeats at 2359 to 2364 and 2386 to 2391. bp 2393 to 2418 are an inverted repeat of 2332 through 2357.

clones were identified and sequenced (Fig. 1). Both LSF cDNAs, representing the same gene, encode all the tryptic peptide sequences that had been obtained from the purified protein (Fig. 1A, boxes; see Discussion). The product of the larger clone is capable of binding two different LSF DNA-binding sites (see below) and encodes a protein with a predicted molecular mass of 57 kDa, similar to the estimated size for LSF of 63 kDa (19). Antipeptide antibodies against two

different tryptic peptides completely supershifted the LSF-DNA complex from whole-cell extracts in a band mobility shift assay (41). Thus, the larger cDNA encodes the biochemical activity previously described as LSF.

The nucleotide sequence of the larger LSF clone was essentially identical, though longer at the 5' end, to the cDNA encoding the human transcription factor α CP2 (23) (the minor discrepancies are noted in the legend to Fig. 1). In addition,

the LSF amino acid sequence is identical to that of LBP-1c (42). Thus, LSF, α CP2, and LBP-1c are identical gene products. The smaller clone, representing LSF-ID, includes all of the coding region of the larger clone except for a 153-bp in-frame deletion removing sequences encoding amino acids 189 to 239 and a 3-bp deletion encoding Gln-491. The LSF-ID amino acid sequence is identical to that of LBP-1d (42). Given that the existing coding regions and the proximal 432 bp of the 5' noncoding regions of the LSF and LSF-ID cDNAs are identical, except for a 1-bp change, the two cDNAs clearly result from expression of a single gene. The divergent region in the 5' untranslated region of LSF-ID cDNA is identical to a portion of an active human transposable element, L1.1 (10). Its significance has not been investigated, although it is intriguing that many alternately spliced mRNAs do utilize alternative promoters (for example, see references 24 and 36).

Indications of important functional regions in proteins can often be gleaned from sequence similarities with other known proteins. A search of the available databases at the National Center for Biotechnology Information using BLASTP (2) revealed, as previously noted (23, 42), that LSF shares extended similarity to only one other known protein, the *Drosophila* transcription factor Elf-1/NTF-1 (7, 11). The *Drosophila* and human proteins are approximately 25% identical over a 427-amino-acid overlap, if the glutamine stretch of LSF, amino acids 404 to 413, is excluded. Not in strict alignment, but present in related regions in both proteins, are multiple copies of the dipeptide repeat serine-proline, which has been implicated in protein-DNA interactions (39). The similarity between Elf-1/NTF-1 and LSF appears to result from similar protein-DNA interactions. Contained within the 427-amino-acid overlap are three regions of high similarity between LSF and Elf-1/NTF-1: amino acids 148 to 159 (66% identity), 205 to 216 (75% identity), and 233 to 246 (79% identity), which were not previously recognized. Significantly, coding sequences for the last two of these three regions are partially or totally missing in LSF-ID (Fig. 1B).

LSF-ID mRNA is present in tissue culture cells. To determine whether the LSF-ID cDNA represented an actual mRNA, total RNA from HeLa and HPB-ALL cells was analyzed by reverse transcription and amplification of an internal region of the RNA. Primers surrounding sequences deleted in LSF-ID were used to amplify coding sequences spanning amino acids 133 to 399. The reaction products were separated by electrophoresis through an agarose gel (Fig. 2). As a positive control, plasmids containing the LSF and LSF-ID cDNAs were amplified together (Fig. 2, lane 2). Two bands corresponding to the expected products from these plasmids were observed. An intermediate band which represents a DNA hybrid composed of one strand of LSF sequences and one strand of LSF-ID sequences was also formed in the reaction. This hybrid also forms if each plasmid is amplified separately and the two reaction mixtures are combined, heated, and cooled (data not shown).

Reverse transcription and amplification of either HeLa (Fig. 2, lane 4) or HPB-ALL (lane 6) RNA resulted in four products. Three of these corresponded to bands observed in the control reaction mixture containing both LSF and LSF-ID plasmid DNAs (Fig. 2, lane 2). The smallest product, RT-LSF, was totally unanticipated. This product has been sequenced and is missing coding sequences for amino acids 306 to 384 (data not shown) and thus represents yet another mRNA isoform. The function of RT-LSF has not yet been investigated. As a crucial control, no products were detected when reverse transcriptase was omitted from the reverse transcription step (Fig. 2, lanes 3 and 5). Thus, clones for both LSF and

FIG. 2. Detection of LSF and LSF-ID mRNAs in HeLa and HPB-ALL cells. An ethidium bromide-stained gel of reverse transcriptase (RT)-PCR products, amplified with primers surrounding the sequences deleted in LSF-ID, is shown. Lanes: 1 (M), DNA markers, with lengths (in base pairs) indicated on the left; 2 (C), control amplification of LSF and LSF-ID cDNA plasmids; 3 and 4, RT-PCR with HeLa RNA with and without RT, respectively; 5 and 6, RT-PCR with HPB-ALL RNA with and without RT, respectively.

LSF-ID represent authentic mRNAs. From this analysis, LSF-ID mRNA and RT-LSF mRNA are at least 15-fold less abundant than the full-length LSF mRNA in both HeLa and HPB-ALL T cells.

LSF-ID is missing an exon present in LSF mRNA. Because both the 5' and the 3' boundaries of the sequences deleted in LSF-ID mRNA are consistent with the boundaries at an intron-exon junction (see Discussion for details), we postulated that the smaller clone was a result of alternative splicing. To test this possibility, genomic DNA obtained from HeLa cells was amplified with combinations of primers either just within or immediately adjacent to the junctions of the deletion: 5' and 3' inner or 5' and 3' outer, respectively (Fig. 3B). When derived from either genomic DNA or the plasmid cDNA, the DNA fragment amplified by the innermost pair of primers was the same length (Fig. 3A, lanes 4 and 12). Thus, the sequences deleted in the LSF-ID cDNA are contained entirely within one exon.

Amplification of genomic DNA with the outermost pair of primers (Fig. 3A, lane 1) resulted in a product of approximately 3,700 bp, as well as two nonspecific bands that do not hybridize to LSF sequences (data not shown). The 3,700-bp product indicates the presence of at least one intron flanking the deleted sequences. Intermediate-size genomic DNA products (approximately 1,700 and 2,100 bp) were amplified with combinations of one inner and one outer primer (Fig. 3A, lanes 2 and 3, respectively), indicating the presence of introns on both sides of the deleted sequences. To verify that all these products represented appropriate amplification of regions in the LSF gene, the amplified DNA fragments were digested with *Pvu*II (Fig. 3A, lanes 5 to 8), which cleaves within the region missing from the LSF-ID cDNA (Fig. 3B). Upon digestion of the 3,700-bp product, the predicted-size products of approximately 1,700 and 2,000 bp appeared (Fig. 3A, lane 5). These products comigrated with the larger DNA products generated by *Pvu*II cleavage of the 1,700- and 2,100-bp amplified DNAs, as expected (Fig. 3A, lanes 6 and 7). The additional 138-bp product that should result from digestion of the 2,100-

FIG. 3. Internal sequences absent in LSF-ID represent a missing exon. Results of PCR amplification of genomic DNA from HeLa cells are shown. (A) Ethidium bromide-stained gel of PCR products. Genomic DNA (lanes 1 to 8) or pET-LSF plasmid DNA (lanes 9 to 16) was amplified with the primers indicated in panel B. A portion of each reaction mixture was digested with *PvuII* (lanes 5 to 8 and lanes 13 to 16). Lanes M contain DNA markers of the indicated sizes (in base pairs). (B) Diagram of predicted genomic structure of the LSF gene surrounding the internal deletion in the LSF-ID cDNA. Exons (boxes), introns (connecting lines), and the region spanning bp 1256 to 1408 (shaded box) (Fig. 1A), the sequences deleted in LSF-ID, are shown. Each of the primer pairs used in PCR and the lanes in panel A which contain the expected products are indicated. The locations of the PCR primers (horizontal arrows) and the *PvuII* site at bp 1384 are indicated. Note that the inner primers contain restriction sites at their 5' termini and thus result in slightly larger DNA products from amplification.

and 170-bp amplified DNAs was also observed (lanes 7 and 8). As size and specificity controls, amplification of the LSF cDNA produced the expected products with all sets of primers, both before and after digestion with *PvuII* (Fig. 3A, lanes 9 to 16). From these experiments, we conclude that the genomic organization of the LSF gene includes an intron of approximately 1,500 bp 5' to the exon missing in LSF-ID and an intron of approximately 1,950 bp 3' to this exon (Fig. 3B).

LSF-ID does not bind to known LSF DNA-binding sites. To characterize the function of LSF-ID, the ability of this protein to interact with known LSF DNA-binding sites was examined

by band mobility shift analysis. Initially, purified fusion proteins composed of GST and either LSF or LSF-ID were incubated with the LSF-280 binding site. As seen in Fig. 4A, GST-LSF effectively bound the LSF-280 site (lanes 2 through 5), as did LSF purified from HeLa cells (lane 9). Neither GST alone (lane 1) nor GST-LSF-ID (lanes 6 through 8) detectably bound any DNA. From these data, we determined that GST-LSF-ID binds DNA at least 1,000-fold less well than GST-LSF. In addition, LSF-ID translated in either wheat germ extract, which lacks endogenous LSF (Fig. 4B, lane 1), or reticulocyte lysate (compare Fig. 6A, lanes 11 and 12) did not detectably bind the LSF-280 site. In these experiments, up to three- to fivefold more radiolabelled LSF-ID protein than radiolabelled LSF was incubated with the DNA-binding site (protein data not shown). Within these limits of detection, in vitro-translated LSF-ID bound DNA at least 50- to 75-fold less well than LSF. Similarly, only LSF, and not LSF-ID, bound the LSF-GC site (data not shown). These data are consistent with the demonstration that in vitro-translated LBP-1d did not bind sequences in the human immunodeficiency virus type 1 long terminal repeat, whereas in vitro-translated LBP-1c did (42).

LSF binds to the LSF-280 site as a homodimer. Given that LSF interacts with directly repeated sequences at one of its high-affinity binding sites, the LSF-GC site (19), and that the LSF consensus site is a direct repeat (22, 32, 42), we predicted that LSF would bind to DNA as a dimer or larger oligomer. The oligomerization state of LSF on the DNA was determined by a modification of the method of Hope and Struhl (18). In their method, different-size derivatives of a protein are incubated together and the electrophoretic mobilities of the resultant mixture of protein-DNA complexes are analyzed. Although LSF derived from the isolated cDNA bound known LSF binding sites, we could not obtain substantially smaller LSF derivatives that could still bind DNA (data not shown). Some DNA binding-competent LSF derivatives, e.g., LSF Δ 24, with an N-terminal deletion of amino acids 1 to 24, did not migrate significantly differently from LSF when bound to DNA (see Fig. 6A); thus, an intermediate complex between full-length LSF and the deleted derivatives could not be observed. We have resolved this dilemma by the use of epitope-specific antibodies.

In our epitope-counting method, an antibody specific to a peptide within a region not required for DNA-binding function was used to supershift the mixture of protein-DNA complexes. The antibody used here recognizes only one of two different LSF species in our analysis. Here, LSF and LSF Δ 24 were cotranslated in wheat germ extract, which lacks endogenous LSF (data not shown). Complexes with LSF-280 DNA were examined by the band mobility shift assay in the presence of antibodies recognizing amino acids 7 to 21. If LSF bound DNA as a dimer, one would expect three distinct protein-DNA complexes: a homodimer of LSF-LSF, a heterodimer of LSF-LSF Δ 24, and a homodimer of LSF Δ 24-LSF Δ 24. Only the electrophoretic mobilities of the first two complexes, LSF-LSF and LSF-LSF Δ 24, would be retarded in the presence of antibody. The first complex could be recognized by two antibodies, while the second could be recognized by only one antibody; thus, the degree of retardation depends on the number of epitopes in the complex (Fig. 5). The complexes formed on radiolabelled DNA with only LSF or only LSF Δ 24 in the presence of antibody are shown in lanes 1 and 6, respectively. As expected, the mobility of the LSF Δ 24-DNA complex was unaltered by antibody, whereas the LSF-DNA complex was supershifted (data not shown). Simultaneous translation of both proteins resulted in one major additional protein-DNA complex of intermediate mobility in the pres-

FIG. 4. LSF-ID does not bind DNA. (A) The following amounts (in nanograms) of the bacterially expressed, purified proteins indicated above the lanes were incubated with LSF-280 DNA: 750 (lane 1), 1.9 (lane 2), 3.8 (lane 3), 7.5 (lane 4), 75 (lane 5), 15 (lane 6), 45 (lane 7), and 90 (lane 8). LSF purified from HeLa cells (lane 9) was used as a positive control. The protein-DNA complexes were resolved on a native polyacrylamide gel. (B) A 1- μ l sample of a wheat germ transcription-translation reaction mixture containing LSF-ID, LSF, or a derivative of LSF that does not bind DNA was incubated with LSF-280 DNA, and the protein-DNA complexes were resolved on a native polyacrylamide gel. The position of the LSF-DNA complex (arrow) is indicated.

ence of antibody. When increasing amounts of plasmid encoding LSF Δ 24 were added in excess to the transcription-translation reaction mixtures in the presence of a fixed amount of LSF plasmid, the intensity of the intermediate band decreased (Fig. 5, lanes 2 to 5). Correspondingly, an increase in the amount of the LSF Δ 24-LSF Δ 24 complex was observed. Because the presence of this antibody stabilizes the interaction of LSF with DNA (data not shown), the ratio of the protein-DNA complexes for LSF-LSF, LSF-LSF Δ 24, and LSF Δ 24-LSF Δ 24 was not 1:2:1, as would be predicted. The simplest interpretation for the formation of three complexes in this experiment is that LSF binds to DNA as a dimer. However, if the antibody required two epitopes in order to supershift the protein-DNA complex, the formation of three types of complexes in the presence of antibody would be consistent with a tetrameric interaction of LSF with DNA as well.

LSF-ID does not affect the binding of LSF to DNA. Because LSF binds DNA as a dimer, we questioned whether LSF-ID would interact with LSF, resulting either in a dominant negative effect on the binding of LSF to DNA or in a protein-DNA complex of altered mobility. To investigate this, increasing amounts of LSF-ID were cotranslated with a fixed amount of LSF or LSF Δ 24 in rabbit reticulocyte lysate, and the resulting DNA-binding activities were analyzed by the band mobility shift assay (Fig. 6). Rabbit reticulocyte lysates were used because they generate higher levels of protein than do wheat germ extracts. Surprisingly, increasing the amount of LSF-ID had no effect on the level of LSF-DNA or LSF Δ 24-DNA complexes (Fig. 6A, lanes 2 to 5 and 7 to 10). The slight decrease (less than twofold) in LSF DNA-binding activity (Fig. 6A, lanes 2 to 5) reflected a decrease in the level of LSF translated in these reactions (Fig. 6B). No decrease in LSF Δ 24

DNA-binding activity was observed (Fig. 6A, lanes 7 to 10); in these reactions, the level of LSF Δ 24 remained relatively constant (Fig. 6B). Reticulocyte lysate contained a small amount of endogenous LSF, whose binding to DNA was also unaffected by an excess of LSF-ID (lanes 11 and 12). Furthermore, LSF-ID did not affect the mobility of the LSF-DNA complex, indicating that it was unlikely to be present in the complex. To support this point, samples containing LSF Δ 24 with increasing levels of LSF-ID (Fig. 6A, lanes 6 to 10) were incubated with the antipeptide antibody that recognizes LSF-ID but not LSF Δ 24 (data not shown). The protein-DNA complex was not supershifted, confirming that LSF-ID was not in the complex. These data suggest that the sequences missing from LSF-ID are likely to be involved in the dimerization of LSF.

The region of LSF missing in LSF-ID contains sequences involved in both DNA-binding function and oligomerization. To investigate further the role of LSF sequences missing in LSF-ID, double amino acid substitutions were made in the two regions of high similarity between LSF and Elf-1/NTF-1 that are missing in LSF-ID (Fig. 1B). Secondary-structure predictions (13) suggest that all three regions of high similarity between the mammalian and *Drosophila* proteins form β -strands (see Discussion) because of alternating hydrophobic and hydrophilic residues. Amino acid substitutions were designed to alter the β -strand structures missing in LSF-ID. Two particular mutants were tested for their ability to bind DNA and to affect the DNA-binding properties of LSF. In LSF234QL/236KE, amino acid Gln-234 was changed to Leu and amino acid Lys-236 was changed to Glu. In LSF211VE/213IR, amino acid Val-211 was changed to Glu and amino acid Ile-213 was changed to Arg.

In a band mobility shift assay, neither mutant protein,

FIG. 5. LSF binds DNA as a dimer. Wheat germ transcription-translation reactions were performed with increasing amounts of pET-LSFΔ24 plasmid in the presence of a constant amount (100 ng) of pET-LSF plasmid (lanes 1 to 5). Lane 6 contained only pET-LSFΔ24 plasmid. pET-11qr, the base vector without LSF sequences, was added as necessary to maintain a fixed amount (500 ng) of total DNA in the reaction mixture. Wheat germ extracts were inoculated as follows: no pET-Δ24 (lane 1); 100 ng of pET-LSFΔ24 (lane 2); 200 ng of pET-LSFΔ24 (lane 3); 300 ng of pET-LSFΔ24 (lane 4); and 400 ng of pET-LSFΔ24 (lanes 5 and 6). A 1-μl sample of each reaction mixture was incubated with radiolabelled LSF-280 DNA and antibody (Ab) and electrophoresed through a native polyacrylamide gel. The mobilities of the homodimer and heterodimer complexes are indicated to the right of the gel. Note that two bands are visible for each dimer in the presence of antibody. Two bands are also observed when complexes of DNA with LSF purified from HeLa cells are incubated with antibody under the same conditions (data not shown). Therefore, the presence of two bands is related to the antibody and not to the form of LSF in the complex.

translated in vitro in rabbit reticulocyte lysates, appreciably bound to DNA (Fig. 7, lanes 6 and 11); the levels of LSF-DNA complexes in these lanes were attributable solely to endogenous LSF levels (lane 12). To confirm this point, neither protein detectably bound DNA when in vitro translated in wheat germ extracts, in the absence of endogenous LSF (data not shown). This inability to bind DNA mimicks that of LSF-ID. Also similar to results for LSF-ID, increasing amounts of the mutant protein LSF211VE/213IR did not affect the DNA-binding activity of LSF, indicating that it cannot act as a dominant negative inhibitor to prevent the binding of LSF to DNA (Fig. 7, compare lane 1 with lanes 7 through 10). Furthermore, LSF211VE/213IR could not be detected in the specific protein-DNA complex, since reaction mixtures containing cotranslated LSF211VE/213IR and LSFΔ24 could not be supershifted by an antibody which recognizes LSF211VE/213IR but not LSFΔ24 (data not shown).

In contrast, the mutant LSF234QL/236KE did reduce the binding of LSF to DNA up to sixfold, after correction for endogenous levels of LSF (Fig. 7, lanes 1 to 5), when the two proteins were cotranslated in reticulocyte lysates. This mutant LSF was also able to inhibit the DNA-binding activity of endogenous LSF in the lysate (Fig. 7, compare lanes 6 and 12). Cotranslation of the wild-type and mutant proteins appears to facilitate the interaction, as the reduction in DNA-binding activity of the endogenous LSF (less than twofold) was not as

FIG. 6. LSF-ID does not affect the LSF-DNA complex. (A) A sample of 100, 200, 300, or 400 ng of pET-LSF-ID DNA was added to either 100 ng of pET-LSF DNA (lanes 2 to 5, respectively) or 100 ng of pET-LSFΔ24 DNA (lanes 7 to 10, respectively) in a rabbit reticulocyte lysate transcription-translation reaction. As a control, reticulocyte lysate was inoculated with 100 ng of pET-LSF DNA (lane 1), 100 ng of pET-LSFΔ24 DNA (lane 6), and 500 ng of pET-LSF-ID DNA (lane 11). Lane 12 displays the small amount of endogenous LSF present in lysate that had been inoculated with 500 ng of pET-11qr, the base expression vector without LSF sequences. pET-11qr was added as necessary to maintain 500 ng of total plasmid DNA in every reaction mixture. A 1-μl sample of each reaction mixture was incubated with radiolabelled LSF-280 DNA and electrophoresed through a native polyacrylamide gel. (B) A 2-μl sample of the reaction mixtures described for panel A was electrophoresed through an SDS-polyacrylamide gel, and the relative amounts of in vitro-translated proteins were quantitated by analysis on the PhosphorImager with ImageQuant software. Levels of either radiolabelled LSF or radiolabelled LSFΔ24 (striped bars) and of radiolabelled LSF-ID (black bars) are indicated. Values were normalized to the amount of radiolabelled LSF in lane 1. Lanes are as in panel A.

substantial as was seen for the newly synthesized LSF. These assays with the mutant LSFs suggest that amino acids 205 to 216 are involved in dimerization.

To examine directly the oligomerization properties of LSF mutants, in vitro protein-protein interactions were assayed with an immobilized GST-LSF fusion protein. GST-LSF, bound to glutathione-Sepharose beads, was subsequently incubated with in vitro-translated LSF, LSFΔ24, LSF-ID, LSF211VE/213IR, or LSF234QL/236KE. The input levels of these labelled proteins are shown in Fig. 8A. After incubation, the beads were washed with a buffer containing either 100 or 500 mM NaCl, and the bound proteins were eluted and analyzed by SDS-polyacrylamide electrophoresis (Fig. 8B and C). A total of 50% of the LSF and LSFΔ24 proteins remained bound following the 100 mM wash, indicating that these

FIG. 7. Effects of amino acid substitutions in LSF on DNA binding and oligomerization. A sample of 100, 200, 300, or 400 ng of pET-LSF234QL/236KE DNA (lanes 2 to 5, respectively) or pET-LSF211VE/213IR DNA (lanes 7 to 10) was added to 100 ng of pET-LSF DNA in a rabbit reticulocyte lysate transcription-translation reaction. Control reaction mixtures were inoculated with 100 ng of pET-LSF DNA (lane 1), 400 ng of pET-LSF234QL/236KE DNA (lane 6), or 400 ng of pET-LSF211VE/213IR DNA (lane 11). Lane 12 displays the low level of endogenous LSF present in lysate that had been inoculated with 500 ng of pET-11qr DNA. pET-11qr was added as necessary to maintain 500 ng of total plasmid DNA in each reaction mixture. A 2- μ l sample of the reaction mixture was incubated with radiolabelled LSF-280 DNA and electrophoresed through a native polyacrylamide gel.

proteins are capable of strong interactions with GST-LSF. Much less protein remained bound with the substitution mutants and LSF-ID (1 to 3%), indicating weaker interactions with GST-LSF. However, LSF234QL/236KE repeatedly bound more tightly than the other two mutants within this region of

FIG. 8. Binding of LSF and mutants to immobilized GST-LSF. Glutathione-Sepharose beads containing GST-LSF were incubated with in vitro-translated proteins, and the bound proteins were examined by SDS-polyacrylamide electrophoresis. (A) Input proteins derived from in vitro transcription-translation in rabbit reticulocyte lysates of 500 ng of pET-LSF Δ 24 DNA (lane 1), pET-LSF DNA (lane 2), pET-LSF234QL/236KE DNA (lane 3), pET-LSF211VE/213IR DNA (lane 4), or pET-LSF-ID DNA (lane 5). (B and C) Proteins bound to GST-LSF beads following four washes with 100 and 500 mM NaCl, respectively. Lanes are as in panel A.

the protein: on average, threefold more than LSF211VE/213IR and 60% more than LSF-ID. In addition, none of these proteins detectably bound to beads containing only GST (data not shown). These results confirm that both LSF and LSF Δ 24 can dimerize with a high affinity and that LSF234QL/236KE can also interact with LSF, albeit with a lower affinity. The data also suggest that LSF211VE/213IR and LSF-ID can interact weakly with LSF; levels in the band mobility shift experiments were presumably not high enough to drive this interaction (Fig. 6A and 7).

LSF234QL/236KE, LSF211VE/213IR, and LSF-ID remained bound to GST-LSF beads to the same extent whether washed with 100 or 500 mM NaCl. However, with the 500 mM NaCl wash, a three- to fourfold reduction in the binding of LSF and LSF Δ 24 to the beads was observed in repeated experiments. This suggests that LSF and LSF Δ 24, which are competent to bind to DNA, can form additional protein-protein contacts that are prevented by mutations in the conserved regions. The salt sensitivity of these LSF-LSF contacts, which may be mediated by β -strands in the conserved regions, implies a hydrophilic interaction. The remaining contacts between the substitution mutants and GST-LSF may be driven by hydrophobic interactions, since they were not affected by an increase in salt concentration.

DISCUSSION

We have isolated two cDNA clones for mammalian transcription factor LSF which reveal its unusual DNA-binding and oligomerization properties. LSF is identical to α CP2 (23) and LBP-1c (42), and LSF-ID, with amino acids 189 to 239 deleted, is identical to LBP-1d (42). While LSF binds to DNA as a dimer, LSF-ID neither binds to DNA nor affects the binding of LSF to DNA. This suggests that the region deleted in LSF-ID overlaps a portion of the protein required for dimerization. Analysis of double point substitutions in subregions conserved between LSF and the *Drosophila* transcription factor Elf-1/NTF-1 and missing in LSF-ID indicates that potential β -strand stretches are important for dimerization.

Characterization of the cDNA clones. Both LSF and LSF-ID contain all six expected tryptic peptide sequences that were obtained from LSF purified from HeLa cells, although Ser-16 contained in one of the peptides was microsequenced as a threonine. Serine is predicted at this position in both clones of LSF as well as the mouse and human clones of α CP2 (m α CP2 and h α CP2, respectively) (23). Other differences between LSF and α CP2 are noted in the legend to Fig. 1.

In agreement with the results presented by Lim et al. (23), we find that LSF/ α CP2 shares extended similarity with only one protein in the database, the *Drosophila* transcription factor Elf-1/NTF-1 (see Results). Strikingly, within the extended region of similarity between LSF and Elf-1/NTF-1, there are three short regions with a high degree of similarity (Fig. 1B). None of the three regions of strong similarity between LSF and Elf-1/NTF-1 is similar to any other protein in the database. Although Bray et al. (7) suggested that Elf-1 may contain a helix-turn-helix DNA-binding motif, on the basis of its weak alignment with MyoD and myogenin and of protein structure predictions, we do not believe that this alignment is significant. Few of the amino acids conserved among Elf-1/NTF-1, MyoD, and myogenin are conserved in LSF. Because Elf-1/NTF-1 is capable of binding the LSF-280 site (32), the sequence similarities between LSF and Elf-1/NTF-1 presumably reflect analogous DNA-binding and dimerization motifs. These may be novel, given that Elf-1/NTF-1 and LSF are unrelated to any other sequence-specific DNA-binding proteins.

LSF contains a novel dimerization motif. Using our epitope-counting method, we have demonstrated that LSF binds to DNA as a dimer. Since LSF interacts with direct repeats in the LSF-GC site (19), and since the LSF consensus sequence is a direct repeat (22, 32, 42), we anticipated that LSF would bind as an oligomer. However, the binding by a dimer was unexpected, as dimeric proteins that bind direct repeats are uncommon (see the introduction). We know that the spacing between the repeats at the LSF-GC site is critical (19). Therefore, in contrast to other dimeric proteins that bind direct repeats, the link between dimerization and DNA-binding surfaces in LSF cannot be highly flexible. A dimeric protein that recognizes only strictly spaced direct repeats in DNA, such as LSF, could bind in a head-to-tail fashion.

We hypothesize that the novelty of the interaction of LSF with DNA may be reflected by an unusual dimerization motif that includes the three regions of greatest similarity between LSF and Elf-1/NTF-1. The first two regions of strongest similarity and the first half of the third region are predicted to form β -strands in both proteins (13). In LSF-ID, which does not readily dimerize with LSF, the second region and part of the third region are deleted. We predict that the β -strands may facilitate inter- or intramolecular protein-protein interactions. Consistent with this hypothesis, NTF-1 also binds DNA as a dimer, requiring a large portion of the protein that includes the region of similarity with LSF (4).

As an initial test of this hypothesis, amino acid substitutions in two regions of similarity between LSF and Elf-1/NTF-1 were constructed. These mutations abolish DNA-binding activity and substantially diminish interactions with LSF. Specifically, LSF234QL/236KE, altered in the third region of similarity, was still capable of weakly interacting with LSF, as assayed both by inhibiting LSF DNA-binding activity and by precipitating with GST-LSF beads. However, LSF211VE/213IR was not capable of inhibiting the DNA-binding activity of LSF, implying a functional role for amino acids 211 and 213 in dimerization.

Interestingly, the binding of LSF-ID, LSF211VE/213IR, and LSF234QL/236KE to GST-LSF beads was unaffected by washing with a high concentration of salt, while the binding of LSF and LSF Δ 24 was reduced three- to fourfold. This implies that the wild-type protein can form additional hydrophilic protein-protein contacts. Two portions of the protein would be required for dimerization of a protein that binds in a head-to-tail fashion. Perhaps the mutants have been altered in only one of these regions, while another region, less sensitive to a high salt concentration, is left intact. An examination of known dimerization motifs in transcription factors, such as the bZIP motif or the helix-loop-helix motif (28), did not reveal any similarities to either primary amino acid sequences or predicted secondary structures in LSF. Thus, this region of LSF required for dimerization is unlike any previously described motifs and may be composed of intermolecular β -sheets mediating ionic interactions with the other LSF monomer. These interpretations regarding the oligomeric interactions of LSF would hold whether LSF binds to DNA as a dimer or, less likely, as a tetramer.

LSF-ID mRNA results from either exon skipping or secondary splicing. Genomic PCR analysis showed that LSF contains an exon that is not present in LSF-ID. The proposed mechanism that is generally described for exon skipping is illustrated in Fig. 9A. An additional mechanism was suggested from an examination of the sequences at the junctions of the deleted exon. These junctions were unexpectedly consistent with sequences within an intron. Note especially the highly conserved GT at the 5' border and AG at the 3' border, almost always found in an intron (Fig. 9B). In addition, the 153-bp exon

A.

B.

FIG. 9. LSF-ID could result from either alternative splicing or secondary splicing of LSF. Exons bordering the deletion in LSF-ID (shaded boxes), the exon that is missing in LSF-ID (hatched box), and introns (lines) are indicated. (A) Alternatively spliced products resulting from exon skipping. (B) Secondary splicing of LSF to create LSF-ID. The consensus sequences of the 5' and 3' splice sites and the consensus sequence for the branch point are indicated above and below the hatched box, respectively. Y and R represent pyrimidines and purines, respectively. Sequences in the LSF cDNA at the appropriate positions and those which agree with the consensus (underlined) are indicated. The branch point (asterisk) is typically found 20 to 40 bp upstream of the 3' splice site. Positions of the sequences are obviously not to scale.

contains a short polypyrimidine track commonly found near a 3' splice site, as well as sequences consistent with the weakly conserved branch point sequence (15, 33). The point of branch formation, almost always an adenine, is rarely found more than 20 to 40 nt upstream of the 3' splice site. In LSF mRNA, an adenine is positioned 34 nt from the exon junction. Finally, although 153 nt is relatively small to be an intron, it is larger than the theoretical minimum of 68 nt (33). Thus, the mRNA for LSF retains all known signals for an intron that could be spliced out to yield LSF-ID mRNA. Due to these compelling sequence alignments, we suggest a novel mechanism whereby the mRNA encoding LSF-ID might also result from secondary splicing of LSF mRNA (Fig. 9B). Examples of sequential splicing have been reported previously. For example, in the preprotachykinin pre-mRNA (26), one intron sequence, intervening sequence 4 (IVS4), must be removed prior to IVS3 or exon 4 will be skipped. In another instance of sequential splicing, the twintron present in the chloroplast gene *psbF* of *Euglena gracilis* (9), an internal intron is removed prior to excision of the external intron. Our model of secondary splicing differs from these examples, as we are not referring to sequential removal of introns but rather to secondary splicing of an exon (intron) from a functional product. To the splicing machinery, the deletion would appear as an exon in the LSF unspliced RNA and as an intron in the LSF mRNA. This would provide an additional level of control of alternative splicing.

Potential biological role of LSF-ID. In our experiments using in vitro-translated protein, LSF-ID did not affect the

binding of LSF to known DNA-binding sites in vitro and was not detected in the LSF-DNA complex. Thus, it did not appear that LSF-ID had a dominant negative effect on the binding of LSF to DNA. In contrast, Yoon et al. (42) demonstrated that with approximately 10-fold larger amounts of bacterially produced LBP-1d (identical to LSF-ID), the DNA-binding activity of in vitro-translated LBP-1 was abolished. We were not able to generate such high levels of LSF-ID in the cotranslation experiments (Fig. 6) to test for a low-affinity interaction between LSF and LSF-ID. However, in the GST-LSF precipitation experiments, a low-affinity interaction was detected, consistent with the data of Yoon et al. (42). Despite the ability of LSF-ID/LBP-1d to inhibit binding of LSF in vitro, the biological relevance of this result remains unclear. The level of LSF-ID mRNA is at least 15-fold lower than that of LSF mRNA in at least two cell types. Furthermore, we have been unable to detect LSF-ID by Western blot analysis of nuclear extracts from several cell types (40). Therefore, the production in vivo of sufficiently high levels of LSF-ID to function as a dominant negative regulator of LSF DNA-binding activity has not been demonstrated.

Perhaps this version of LSF, resulting from alternative splicing, has regulatory capabilities different from those of LSF. First, it is possible that LSF-ID recognizes other DNA sequences which differ from the LSF-280, LSF-GC, and human immunodeficiency virus long terminal repeat sites. Variants of the *Drosophila* chorion transcription factor, CF2-I and CF2-II, and alternatively spliced products of the Wilms' tumor gene, *wt1*, bind different, though related, DNA sequences (6, 14). Second, LSF-ID may regulate the transcriptional activation function, rather than DNA-binding activity, of LSF. That is, it may interact with an accessory protein that is required for LSF to regulate transcription in vivo, perhaps in a dominant negative manner. Finally, even though LSF-ID mRNA is much less abundant than LSF mRNA in HeLa and HPB-ALL cell lines, LSF-ID mRNA may be upregulated in certain cell types in animals or during development. Under such circumstances, LSF-ID could play a prominent role in altering transcriptional regulation by LSF.

ACKNOWLEDGMENTS

We thank Wendy Weller for assistance in two-dimensional protein gel electrophoresis, William S. Lane at the Harvard Microchemistry Facility for the tryptic peptide analysis, Brian Seed for the generous gift of a HPB-ALL T-cell plasmid library, Christina Powell for assistance in DNA sequencing, and William G. Kaelin for the pGEX-2TK plasmid. We thank Wendy Hanna-Rose for critically reading the manuscript.

This research was supported by grants CA38038 (U.H.) and CA45285 (D.C.P.) awarded by the National Cancer Institute. M.S. was supported in part by the National Science Foundation.

REFERENCES

- Abersold, R. H., J. Leavitt, R. A. Saavedra, L. E. Hood, and S. B. H. Kent. 1987. Internal amino acid sequence analysis of proteins separated by one- or two-dimensional gel electrophoresis after *in situ* protease digestion on nitrocellulose. *Proc. Natl. Acad. Sci. USA* **84**:6970-6974.
- Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**:403-410.
- Aruffo, A., and B. Seed. 1987. Molecular cloning of a CD28 cDNA by a high-efficiency COS cell expression system. *Proc. Natl. Acad. Sci. USA* **84**:8573-8577.
- Attardi, L. D., and R. Tjian. 1993. *Drosophila* tissue-specific transcription factor NTF-1 contains a novel isoleucine rich activation motif. *Genes Dev.* **7**:1341-1353.
- Ausubel, F. M., R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith, and K. Struhl. 1991. Current protocols in molecular biology. Greene Publishing Associates/Wiley Interscience, New York.
- Bickmore, W. A., K. Oghene, M. H. Little, A. Seawright, and V. vanHeyningen. 1992. Modulation of DNA binding specificity by alternative splicing of the Wilms tumor *wt1* gene transcript. *Science* **257**:235-237.
- Bray, S. J., B. Burke, N. H. Brown, and J. Hirsh. 1989. Embryonic expression pattern of a family of *Drosophila* proteins that interact with a central nervous system regulatory element. *Genes Dev.* **3**:1130-1145.
- Cooney, A. J., S. Y. Tsai, B. W. O'Malley, and M.-J. Tsai. 1992. Chicken ovalbumin upstream promoter transcription factor (COUP-TF) dimers bind to different GGTCA response elements, allowing COUP-TF to repress hormonal induction of the vitamin D₃, thyroid hormone, and retinoic acid receptors. *Mol. Cell. Biol.* **12**:4153-4163.
- Copertino, D. W., and R. B. Hallick. 1991. Group II twintron: an intron within an intron in a chloroplast cytochrome b-559 gene. *EMBO J.* **10**:433-442.
- Dombroski, B. A., S. L. Mathias, E. Nanthakumar, A. F. Scott, and J. H. H. Kazazian. 1991. Isolation of an active human transposable element. *Science* **254**:1805-1808.
- Dynlacht, B. D., L. D. Attardi, A. Admon, M. Freeman, and R. Tjian. 1989. Functional analysis of NTF-1, a developmentally regulated *Drosophila* transcription factor that binds neuronal *cis* elements. *Genes Dev.* **3**:1677-1688.
- Freedman, L. P., and T. L. Towers. 1991. DNA binding properties of the vitamin D₃ receptor zinc finger region. *Mol. Endocrinol.* **5**:1815-1826.
- Garnier, J., D. J. Osguthorpe, and B. Robson. 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120**:97-120.
- Gogos, J. A., T. Hsu, J. Bolton, and F. C. Kafatos. 1992. Sequence discrimination by alternatively spliced isoforms of a DNA binding zinc finger domain. *Science* **257**:1951-1955.
- Green, M. R. 1986. Pre-mRNA splicing. *Annu. Rev. Genet.* **20**:671-708.
- Higuchi, R. 1990. Recombinant PCR, p. 177-183. In M. A. Innis, D. H. Gelfand, J. J. Sninsky, and T. J. White (ed.), *PCR protocols: a guide to methods and applications*. Academic Press, Inc., San Diego, Calif.
- Ho, Y. S., M. E. Mohoney, D. L. Wulff, and M. Rosenberg. 1988. Identification of the DNA binding domain of the phage lambda cII transcriptional activator and the direct correlation of cII protein stability with its oligomeric forms. *Genes Dev.* **2**:184-195.
- Hope, I. A., and K. Struhl. 1987. GCN4, a eukaryotic transcriptional activator protein, binds as a dimer to target DNA. *EMBO J.* **6**:2781-2784.
- Huang, H.-C., R. Sundseth, and U. Hansen. 1990. Transcription factor LSF binds two variant bipartite sites within the SV40 late promoter. *Genes Dev.* **4**:287-298.
- Kaelin, W. G., D. C. Pallas, J. A. DeCaprio, F. J. Kaye, and D. M. Livingston. 1991. Identification of cellular proteins that can interact specifically with the T/E1A-binding region of the Retinoblastoma gene product. *Cell* **64**:521-532.
- Kim, C. H., C. Heath, A. Bertuch, and U. Hansen. 1987. Specific stimulation of simian virus 40 late transcription *in vitro* by a cellular factor binding the simian virus 40 21-base-pair repeat promoter element. *Proc. Natl. Acad. Sci. USA* **84**:6025-6029.
- Lim, L. C., L. Fang, S. L. Swendeman, and M. Sheffery. 1993. Characterization of the molecularly cloned murine α -globin transcription factor CP2. *J. Biol. Chem.* **268**:18008-18017.
- Lim, L. C., S. L. Swendeman, and M. Sheffery. 1992. Molecular cloning of the α -globin transcription factor CP2. *Mol. Cell. Biol.* **12**:828-835.
- Link, D. C., S. J. Gutkind, K. C. Robbins, and T. J. Ley. 1992. Characterization of the 5' untranslated region of the human *c-fgr* gene and identification of the major myelomonocytic *c-fgr* promoter. *Oncogene* **7**:877-884.
- Manley, J. L., A. Fire, A. Cano, P. A. Sharp, and M. L. Gelfand. 1980. DNA-dependent transcription of adenovirus genes in a soluble whole-cell extract. *Proc. Natl. Acad. Sci. USA* **77**:3855-3859.

26. Nasim, F. H., P. A. Spears, H. M. Hoffmann, H.-C. Kuo, and P. J. Grabowski. 1990. A sequential splicing mechanism promotes selection of an optional exon by repositioning a downstream 5' splice site in preprotachykinin pre-mRNA. *Genes Dev.* **4**:1172-1184.
27. O'Farrell, P. H. 1975. High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250**:4007-4021.
28. Pabo, C. O., and R. T. Sauer. 1992. Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.* **61**:1053-1095.
29. Perisic, O., H. Xiao, and J. T. Lis. 1989. Stable binding of *Drosophila* heat shock factor to head-to-head and tail-to-tail repeats of a conserved 5 bp recognition unit. *Cell* **59**:797-806.
30. Sauer, R. T., D. L. Smith, and A. D. Johnson. 1988. Flexibility of the yeast $\alpha 2$ repressor enables it to occupy the ends of its operator, leaving the center free. *Genes Dev.* **2**:807-816.
31. Seed, B., and A. Aruffo. 1987. Molecular cloning of the CD2 antigen, the T-cell erythrocyte receptor, by a rapid immunoselection procedure. *Proc. Natl. Acad. Sci. USA* **84**:3365-3369.
32. Shirra, M., and U. Hansen. Unpublished data.
33. Smith, C. W. J., J. Patton, and B. Nadal-Ginard. 1989. Alternative splicing in the control of gene expression. *Annu. Rev. Genet.* **23**:527-577.
34. Smith, D. L., and A. D. Johnson. 1992. A molecular mechanism for combinatorial control in yeast: MCM1 protein sets the spacing and orientation of the homeodomains of an $\alpha 2$ dimer. *Cell* **68**:133-142.
35. Sorger, P. K., and H. C. Nelson. 1989. Trimerization of a yeast transcriptional activator via a coiled-coil motif. *Cell* **59**:807-813.
36. Spohr, G. 1991. *Xenopus laevis* c-myc I and II genes: molecular structure and developmental expression. *Nucleic Acids Res.* **19**:3081-3088.
37. Stone, K., M. B. LoPresti, N. D. Williams, J. M. Crawford, R. DeAngelis, and K. R. Williams. 1989. Enzyme digestion of proteins and HPLC peptide isolation in subnanomole range, p. 377-391. In T. E. Hugli (ed.), *Techniques in protein chemistry*. Academic Press, New York.
38. Sundseth, R., and U. Hansen. 1992. Activation of RNA polymerase II transcription by the specific DNA-binding protein LSF. *J. Biol. Chem.* **267**:7845-7855.
39. Suzuki, M. 1989. SPXX, a frequent sequence motif in gene regulatory proteins. *J. Mol. Biol.* **207**:61-84.
40. Volker, J. L., Q. Zhu, and U. Hansen. Unpublished data.
41. Volker, J. L., L. Rameh, J. DeCaprio, and U. Hansen. Submitted for publication.
42. Yoon, J.-B., G. Li, and R. G. Roeder. 1994. Characterization of a family of related cellular transcription factors which can modulate human immunodeficiency virus type 1 transcription in vitro. *Mol. Cell. Biol.* **14**:1776-1785.