The B-cell-specific Oct-2 protein contains POU box- and homeo box-type domains

Roger G. Clerc,[1] Lynn M. Corcoran,[3] Jonathan H. LeBowitz,[1] David Baltimore,[2,3] and Phillip A. Sharp[1,2]

[1]Center for Cancer Research and [2]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 USA; [3]Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142 USA

Transcription of promoters of immunoglobulin genes is controlled by an octanucleotide sequence element. The sequence of a cDNA encoding a B-cell-specific protein, Oct-2, has been determined. This protein specifically recognizes the octanucleotide element and is part of the previously identified NF-A2 family of proteins. The DNA-binding domain of Oct-2 is structurally related to the homeo box consensus and thus contains a potential helix-turn-helix sequence. Oct-2 also possesses a potential 'leucine zipper' domain, where four leucines are each separated by exactly seven residues. Comparisons of Oct-2 with protein Oct-1, which also recognizes the octanucleotide element but is constitutively expressed in all cell types, show high sequence conservation through the 60-residue DNA-binding domain, as well as an adjacent tract of 75 residues. The latter conserved region is also found in regulatory genes expressed in pituitary cells and nematodes and has been termed a POU box. Because two different cDNAs were isolated, it is proposed that the oct-2 gene is expressed as multiple mRNAs that vary in splicing patterns. Most interestingly, the oct-2 cDNA contains a second overlapping open reading frame, 278 residues in length, which might also specify a protein important for B-cell development.

Expression of immunoglobulin genes is restricted to B lymphocytes and is dependent upon several cis-acting elements (Banerji et al. 1983; Gillies et al. 1983; Grosschedl and Baltimore 1985; Gerster et al. 1987; Peterson and Calame 1987). Both the promoter and the enhancer of immunoglobulin genes direct tissue-specific transcription, and in the case of the heavy-chain gene, both contain the well-conserved octanucleotide sequence (ATGCAAAT) (Falkner and Zachau 1984; Parslow et al. 1984; Eaton and Calame 1987). In the light-chain gene, this octanucleotide sequence is found only in the promoter. Analysis of mutations that alter the octanucleotide element suggests that it is responsible primarily for the B-cell-specific activation of immunoglobulin genes, (Bergman et al. 1984; Mason et al. 1985). Furthermore, insertion of the octanucleotide element into the β-globin promoter is sufficient to confer B-cell-specific activation on this heterologous promoter (Wirth et al. 1987).

The same octanucleotide sequence found in immunoglobulin regulatory elements is also present in promoters that are expressed in all cell types. Mutational analysis has shown that this sequence is important for transcription of the small nuclear (sn) RNA genes (Carbon et al. 1987) and the histone H2B gene (LaBella et al. 1988) in B cells and non-B cells.

Two proteins that specifically bind the octanucleotide element have been described using a gel electrophoresis assay for DNA–protein complexes. The nuclear factor NF-A1, which is constitutively expressed in all cell types, generates a slower migrating complex with an octanucleotide probe than the nuclear factor NF-A2, whose presence is restricted to B cells (Singh et al. 1986; Staudt et al. 1986). Proteins with similar properties have been purified by affinity chromatography. The constitutive factor NF-A1 (OTF-1 or OBP100) has a molecular weight of ~100kD (Fletcher et al. 1987; Sturm et al. 1987). This factor is probably identical to NF-III, a cellular factor required for adenovirus replication (O'Neill and Kelly 1988; O'Neill et al. 1988). The B-cell-specific factor NF-A2 (OTF-2) has a molecular mass of 59–63 kD (Scheidereit et al. 1987, and see below).

It was suggested that NF-A2 was responsible for the B-cell-specific activity of the octanucleotide element in immunoglobulin promoters. Similarly, NF-A1 was projected as the transcription factor that specifically recognized the octanucleotide element in all cell types (Staudt et al. 1986). This situation defines a paradox: Why is the constitutive factor NF-A1 capable of stimulating transcription of some promoters in non-B cells but is incapable of stimulating high levels of transcription of the immunoglobulin promoters in these cells (LeBowitz et al. 1988)?

We have addressed this problem previously by using a novel technology (Singh et al. 1988) to isolate a partial cDNA segment encoding an octanucleotide-binding pro-

tein (Staudt et al. 1988). A cDNA library was constructed in the expression phage λgt11 using poly(A)⁺ RNA from the human B-cell line BJAB. This library was screened for expression of a β-galactosidase fusion protein that specifically bound a DNA probe consisting of an oligomer of the octanucleotide element. A specific cDNA, derived from the oct-2 gene, was isolated and shown to encode a protein that binds the octanucleotide sequence with the same specificity as NF-A2 and to hybridize to human DNA as a single-copy gene (Staudt et al. 1988). Most importantly, the oct-2 cDNA hybridized to poly(A)⁺ RNA from B cells and not to equivalent RNA samples from non-B cells. This suggested that the oct-2 gene encoded NF-A2, the B-cell-specific transcription factor. Interestingly, oct-2 hybridized to a number of mRNAs in B cells with lengths from 1 to 7 kb. The nature of the length variations among these mRNAs was not determined.

In this paper we report on the characterization of a cDNA, oct-2, encoding the B-cell-specific factor NF-A2. This protein will be referred to as the Oct-2 factor. In addition, a second cDNA for NF-A2 specifying an alternate carboxy-terminal domain is described.

Results

Nucleotide and predicted amino acid sequences: a family of B-cell-specific octanucleotide-binding proteins

The oct-2 cDNA segment (1.1 kb) of phage 3-1, which was isolated originally from a λgt11 expression library generated with RNA from the human B-cell line BJAB, was used to identify additional recombinants in the same library. One of these phage (pass-3) contained a 1.8-kb DNA insert. Sequence analysis of the cDNA segment in the original λgt11 phage (3-1) revealed a long open reading frame (ORF) that was ended with multiple nonsense codons at its 3′ terminus. Sequence analysis of the pass-3 segment yielded an identical sequence through the ORF, but an abrupt transition to a novel sequence occurred at the carboxyl terminus (Fig. 1B; see below). The amino terminus of the ORF in both of these cDNA segments was not represented in the cDNA inserts. Additional recombinants from the λgt11 library were identified by screening with a probe from the amino-terminal portion of the pass-3 segment. This resulted in the isolation of a 0.75-kb cDNA segment (pass 5.5) whose sequence extended the amino-terminal portion of the previously identified ORF. In this cDNA segment, a nonsense codon is found 36 bp upstream of the first AUG in the open frame. The sequence context of this AUG conforms well to that expected for an initiation codon (Kozak 1986). Two other AUG codons occur at positions 6 and 13 in the reading frame. Each of these also has an excellent context for initiation. The amino terminus of the protein has been assigned arbitrarily to the 5′-most AUG codon. The cDNA sequence extends 66 bases, 5′ from this position, but the total length of the 5′-untranslated region has not been determined.

The sequences of pass-5.5, pass-3, and 3-1 were combined to form an ORF encoding a protein of 467 amino acids in length (Fig. 1A). Data presented below suggest that this ORF encodes one form of NF-A2 (Oct-2). The amino acid sequence of Oct-2 has several interesting features (Fig. 1C). It contains three glutamine (Q)-rich blocks (ranging from 50% of Q content) in the amino-terminal part of the polypeptide, beginning at nucleotide positions 376, 448, and 502, and a comparably acidic region [aspartic (E) or glutamic (D) amino acids] between positions 648 and 678. Clusters of Q residues, as well as E or D amino acids, have been described previously in many transcription factors, and there is strong evidence that acidic regions function in the activation of RNA polymerase II transcription initiation. The region of Oct-2 responsible for sequence-specific DNA binding (DNA) is discussed below. Downstream of this position is a series of four leucine residues separated by exactly seven amino acids (positions 1231–1294 in Fig. 1A). A similar configuration of leucine residues in the transcription factor C/EBP has been suggested to form an amphipathic α-helical structure, where the leucine residues are arranged along one side of the helix. Two such helices are thought to interact by a 'leucine zipper' mechanism generating a dimeric protein (Landschultz et al. 1988a,b).

Searches for sequence similarities in the GenBank library revealed that a region of the Oct-2 protein from position 955 to 1134 was related distantly to a family of proteins containing homeo boxes (Gehring 1987). Figure 6 shows that a 60-amino-acid region of Oct-2 contains

Figure 1. (See page 1572.) (A) Nucleotide sequence of oct-2 gene derived from cDNA and predicted amino acid sequence of encoded proteins. Amino acid sequence of Oct-2 protein depicted in plain uppercase letters. cDNA-clone pass-5.5 spans from position 1 (5′ end) to position 750 (3′ end). cDNA clone pass-3, 5′ end and 3′ end, are, respectively, at position 92 here and 1850 in B. cDNA clone 3-1 starts at position 655 and ends at position 1717. The nucleotide sequence shown in A was reconstructed by merging the DNA sequences from clone pass-5.5 from position 1 to 100, from clone pass-3 from 100 to 660, and from clone 3-1 from position 660 to 1717. Extensive nucleotide sequence overlaps were available to allow unequivocal merges. Sequence of protein encoded by the long overlapping open reading frame (LORF, 278 amino acids) is also shown in uppercase letters starting at nucleotide position 735. Wavy arrows delimit the glutamine (Q)-rich, glutamic- and aspartic (E/D)-rich, and glycine (G)-rich regions, respectively. Solid arrows delimit the helix-turn-helix motif. Boxed leucine (L) residues are spaced exactly by seven residues. The vertical arrow indicates the position where the nucleotide sequence diverges with that shown in B. Stars indicate stop codons. (B) Nucleotide sequence of the 3′ terminus and predicted amino acid sequence of the carboxyl terminus derived from clone pass-3. The code is the same as in A, except the LORF is shown in italic letters. The vertical arrow denotes the divergence point of the DNA sequence. (C) Schematic representation of the amino acid sequence deduced from oct-2 gene-derived cDNA. The code is as in A. The DNA-binding domain is depicted as DNA, and the region containing the four regularly spaced L residues is boxed in. (N) Amino terminus; (C) carboxyl terminus.
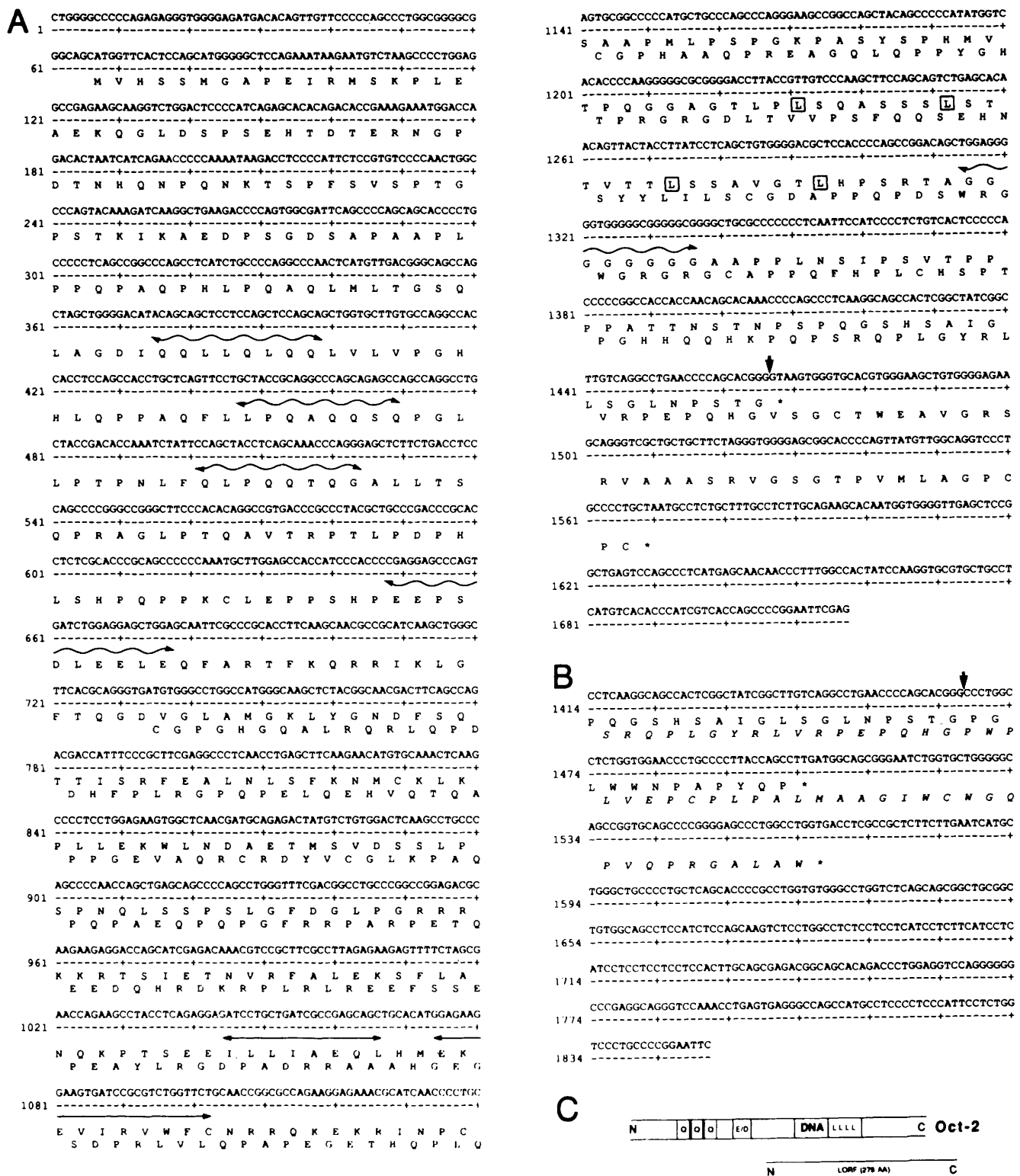
A

```
      CTGGGGCCCCCAGAGAGGGTGGGGAGATGACACAGTTGTTCCCCCAGCCCTGGCGGGGCG
   1  ---------+---------+---------+---------+---------+---------+

      GGCAGCATGGTTCACTCCAGCATGGGGGCTCCAGAAATAAGAATGTCTAAGCCCCTGGAG
  61  ---------+---------+---------+---------+---------+---------+
             M  V  H  S  S  M  G  A  P  E  I  R  M  S  K  P  L  E

      GCCGAGAAGCAAGGTCTGGACTCCCCATCAGAGCACACAGACACCGAAAGAAATGGACCA
 121  ---------+---------+---------+---------+---------+---------+
       A  E  K  Q  G  L  D  S  P  S  E  H  T  D  T  E  R  N  G  P

      GACACTAATCATCAGAACCCCCAAAATAAGACCTCCCCATTCTCCGTGTCCCAACTGGC
 181  ---------+---------+---------+---------+---------+---------+
       D  T  N  H  Q  N  P  Q  N  K  T  S  P  F  S  V  S  P  T  G

      CCCAGTACAAAGATCAAGGCTGAAGACCCCAGTGGCGATTCAGCCCCAGCAGCACCCCTG
 241  ---------+---------+---------+---------+---------+---------+
       P  S  T  K  I  K  A  E  D  P  S  G  D  S  A  P  A  A  P  L

      CCCCCTCAGCCGGCCCAGCCTCATCTGCCCCAGGCCCAACTCATGTTGACGGGCAGCCAG
 301  ---------+---------+---------+---------+---------+---------+
       P  P  Q  P  A  Q  P  H  L  P  Q  A  Q  L  M  L  T  G  S  Q

      CTAGCTGGGGACATACAGCAGCTCCTCCAGCTCCAGCAGCTGGTGCTTGTGCCAGGCCAC
 361  ---------+---------+---------+---------+---------+---------+
       L  A  G  D  I  Q  Q  L  L  Q  L  Q  Q  L  V  L  V  P  G  H

      CACCTCCAGCCACCTGCTCAGTTCCTGCTACCGCAGGCCCAGCAGAGCCAGCCAGGCCTG
 421  ---------+---------+---------+---------+---------+---------+
       H  L  Q  P  P  A  Q  F  L  L  P  Q  A  Q  Q  S  Q  P  G  L

      CTACCGACACCAAATCTATTCCAGCTACCTCAGCAAACCCAGGGAGCTCTTCTGACCTCC
 481  ---------+---------+---------+---------+---------+---------+
       L  P  T  P  N  L  F  Q  L  P  Q  Q  T  Q  G  A  L  L  T  S

      CAGCCCCGGGCCGGGCTTCCCACACAGGCCGTGACCCGCCCTACGCTGCCCGACCCGCAC
 541  ---------+---------+---------+---------+---------+---------+
       Q  P  R  A  G  L  P  T  Q  A  V  T  R  P  T  L  P  D  P  H

      CTCTCGCACCCGCAGCCCCCCAAATGCTTGGAGCCACCATCCCACCCCGAGGAGCCCAGT
 601  ---------+---------+---------+---------+---------+---------+
       L  S  H  P  Q  P  P  K  C  L  E  P  P  S  H  P  E  E  P  S

      GATCTGGAGGAGCTGGAGCAATTCGCCCGCACCTTCAAGCAACGCCGCATCAAGCTGGGC
 661  ---------+---------+---------+---------+---------+---------+
       D  L  E  E  L  E  Q  F  A  R  T  F  K  Q  R  R  I  K  L  G

      TTCACGCAGGGTGATGTGGGCCTGGCCATGGGCAAGCTCTACGGCAACGACTTCAGCCAG
 721  ---------+---------+---------+---------+---------+---------+
       F  T  Q  G  D  V  G  L  A  M  G  K  L  Y  G  N  D  F  S  Q
                   C  G  P  H  G  Q  A  L  R  Q  R  L  Q  P  D

      ACGACCATTTCCCGCTTCGAGGCCCTCAACCTGAGCTTCAAGAACATGTGCAAACTCAAG
 781  ---------+---------+---------+---------+---------+---------+
       T  T  I  S  R  F  E  A  L  N  L  S  F  K  N  M  C  K  L  K
       D  H  F  P  L  R  G  P  Q  P  E  L  Q  E  H  V  Q  T  Q  A

      CCCCTCCTGGAGAAGTGGCTCAACGATGCAGAGACTATGTCTGTGGACTCAAGCCTGCCC
 841  ---------+---------+---------+---------+---------+---------+
       P  L  L  E  K  W  L  N  D  A  E  T  M  S  V  D  S  S  L  P
       P  P  G  E  V  A  Q  R  C  R  D  Y  V  C  G  L  K  P  A  Q

      AGCCCCAACCAGCTGAGCAGCCCCAGCCTGGGTTTCGACGGCCTGCCCGGCCGGAGACGC
 901  ---------+---------+---------+---------+---------+---------+
       S  P  N  Q  L  S  S  P  S  L  G  F  D  G  L  P  G  R  R  R
       P  Q  P  A  E  Q  P  Q  P  G  F  R  R  P  A  R  P  E  T  Q

      AAGAAGAGGACCAGCATCGAGACAAACGTCCGCTTCGCCTTAGAGAAGAGTTTTCTAGCG
 961  ---------+---------+---------+---------+---------+---------+
       K  K  R  T  S  I  E  T  N  V  R  F  A  L  E  K  S  F  L  A
       E  E  D  Q  H  R  D  K  R  P  L  R  L  R  E  E  F  S  S  E

      AACCAGAAGCCTACCTCAGAGGAGATCCTGCTGATCGCCGAGCAGCTGCACATGGAGAAG
1021  ---------+---------+---------+---------+---------+---------+
       N  Q  K  P  T  S  E  E  I  L  L  I  A  E  Q  L  H  M  E  K
       P  E  A  Y  L  R  G  D  P  A  D  R  R  A  A  A  H  G  E  G

      GAAGTGATCCGCGTCTGGTTCTGCAACCGGCGCCAGAAGGAGAAACGCATCAACCCCTGC
1081  ---------+---------+---------+---------+---------+---------+
       E  V  I  R  V  W  F  C  N  R  R  Q  K  E  K  R  I  N  P  C
       S  D  P  R  L  V  L  Q  P  A  P  E  G  E  T  H  Q  P  L  Q
```

```
      AGTGCGGCCCCCATGCTGCCCAGCCCCAGGGAAGCCGGCCAGCTACAGCCCCCATATGGTC
1141  ---------+---------+---------+---------+---------+---------+
       S  A  A  P  M  L  P  S  P  G  K  P  A  S  Y  S  P  H  M  V
       C  G  P  H  A  A  Q  P  R  E  A  G  Q  L  Q  P  P  Y  G  H

      ACACCCCAAGGGGGCGCGGGGACCTTACCGTTGTCCCAAGCTTCCAGCAGTCTGAGCACA
1201  ---------+---------+---------+---------+---------+---------+
       T  P  Q  G  G  A  G  T  L  P [L] S  Q  A  S  S  S [L] S  T
       T  P  R  G  R  G  D  L  T  V  V  P  S  F  Q  Q  S  E  H  N

      ACAGTTACTACCTTATCCTCAGCTGTGGGGACGCTCCACCCCAGCCGGACAGCTGGAGGG
1261  ---------+---------+---------+---------+---------+---------+
       T  V  T  T [L] S  S  A  V  G  T [L] H  P  S  R  T  A  G  G
       S  Y  Y  L  I  L  S  C  G  D  A  P  P  Q  P  D  S  W  R  G

      GGTGGGGGCGGGGGCGGGGCTGCGCCCCCCCTCAATTCCATCCCCTCTGTCACTCCCCCA
1321  ---------+---------+---------+---------+---------+---------+
       G  G  G  G  G  A  A  P  P  L  N  S  I  P  S  V  T  P  P
       W  G  R  G  R  G  C  A  P  P  Q  F  H  P  L  C  H  S  P  T

      CCCCCGGCCACCACCAACAGCACAAACCCCAGCCCTCAAGGCAGCCACTCGGCTATCGGC
1381  ---------+---------+---------+---------+---------+---------+
       P  P  A  T  T  N  S  T  N  P  S  P  Q  G  S  H  S  A  I  G
       P  G  H  H  Q  Q  H  K  P  Q  P  S  R  Q  P  L  G  Y  R  L

      TTGTCAGGCCTGAACCCCAGCACGGGGTAAGTGGGTGCACGTGGGAAGCTGTGGGGAGAA
1441  ---------+---------+---------+---------+---------+---------+
       L  S  G  L  N  P  S  T  G  *
       V  R  P  E  P  Q  H  G  V  S  G  C  T  W  E  A  V  G  R  S

      GCAGGGTCGCTGCTGCTTCTAGGGTGGGGAGCGGCACCCCAGTTATGTTGGCAGGTCCCT
1501  ---------+---------+---------+---------+---------+---------+
          R  V  A  A  A  S  R  V  G  S  G  T  P  V  M  L  A  G  P  C

      GCCCCTGCTAATGCCTCTGCTTTGCCTCTTGCAGAAGCACAATGGTGGGGTTGAGCTCCG
1561  ---------+---------+---------+---------+---------+---------+
          P  C  *

      GCTGAGTCCAGCCCTCATGAGCAACAACCCTTTGGCCACTATCCAAGGTGCGTGCTGCCT
1621  ---------+---------+---------+---------+---------+---------+

      CATGTCACACCCATCGTCACCAGCCCCGGAATTCGAG
1681  ---------+---------+---------+-------
```

B

```
      CCTCAAGGCAGCCACTCGGCTATCGGCTTGTCAGGCCTGAACCCCAGCACGGGCCCTGGC
1414  ---------+---------+---------+---------+---------+---------+
       P  Q  G  S  H  S  A  I  G  L  S  G  L  N  P  S  T  G  P  G
       S  R  Q  P  L  G  Y  R  L  V  R  P  E  P  Q  H  G  P  W  P

      CTCTGGTGGAACCCTGCCCCTTACCAGCCTTGATGGCAGCGGGAATCTGGTGCTGGGGGC
1474  ---------+---------+---------+---------+---------+---------+
       L  W  W  N  P  A  P  Y  Q  P  *
       L  V  E  P  C  P  L  P  A  L  M  A  A  G  I  W  C  W  G  Q

      AGCCGGTGCAGCCCCGGGGAGCCCTGGCCTGGTGACCTCGCCGCTCTTCTTGAATCATGC
1534  ---------+---------+---------+---------+---------+---------+
       P  V  Q  P  R  G  A  L  A  W  *

      TGGGCTGCCCCTGCTCAGCACCCCGCCTGGTGTGGGCCTGGTCTCAGCAGCGGCTGCGGC
1594  ---------+---------+---------+---------+---------+---------+

      TGTGGCAGCCTCCATCTCCAGCAAGTCTCCTGGCCTCTCCTCCTCATCCTCTTCATCCTC
1654  ---------+---------+---------+---------+---------+---------+

      ATCCTCCTCCTCCTCCACTTGCAGCGAGACGGCAGCACAGACCCTGGAGGTCCAGGGGGG
1714  ---------+---------+---------+---------+---------+---------+

      CCCGAGGCAGGGTCCAAACCTGAGTGAGGGCCAGCCATGCCTCCCCTCCCATTCCTCTGG
1774  ---------+---------+---------+---------+---------+---------+

      TCCCTGCCCCGGAATTC
1834  ---------+-------
```

C

N [o o o] [E/D] [DNA] [LLLL] C   **Oct-2**

N   LORF (275 AA)   C

Figure 1. *(See preceding page for legend.)*

30% identity with the prototype homeo box sequence in the *Antennapedia* (*Antp*) protein. That the homeo box specifies a sequence-specific DNA-binding domain is argued most strongly by its homology with the DNA-binding domain of the yeast mating regulatory protein, MATα (Astell et al. 1981; Scott and Weiner 1984). The homologous regions in these proteins can be folded into a helix-turn-helix structure similar to that first identified in the structural analysis of phage λ repressor (for a review, see Pabo and Sauer 1984). A prediction of the most probable secondary structure of Oct-2 also revealed a helix-turn-helix structure between the residues of isoleucine (position 1045) and cysteine (position 1102). This, by analogy, we propose that this region of Oct-2 specifies the sequence-specific binding of the protein.

As mentioned above, sequences at the 3' end of the pass-3 recombinant abruptly diverged from that of recombinant 3-1 at the position (1466) of its termination codon (see vertical arrow in Fig. 1B). The substituted sequences in the second recombinant, pass-3, extended the reading frame of the Oct-2-related protein by an additional 12 amino acids. To rule out a possible artifactual sequence generated by the insertion of a fragment during construction of the cDNA library, total poly(A)+ RNA

from the BJAB cell line was analyzed by Northern blot with a DNA fragment from the novel 3'-terminal portion of the pass-3 cDNA. In Figure 2A, lane 2, this specific probe hybridized only to the two fastest-migrating mRNAs of the total family of six mRNAs, which were detected by hybridization with the total 3-1 cDNA (Fig. 2A, lane 1). A similar specific probe was excised from the 3' terminus of the 3-1 cDNA. In constrast, this probe only hybridized to the two slowest-migrating mRNAs in the total family of six (Fig. 2A, lane 3). This suggests that the two cDNA segments correspond to different populations of oct-2 mRNAs. Still other 3'-terminal sequences must be present on oct-2 mRNAs, as several species did not hybridize with either cDNA-specific probe.

The proteins encoded by the two cDNAs should only differ at their carboxyl termini by 12 amino acids or ~1.5 kD. In vitro transcription/translation of subfragments of the 3-1 and pass-3 recombinants was used to confirm this prediction. Fragments representing the 3' portions of 3-1 and pass-3 were subcloned into the expression plasmid pBS-ATG (see legend to Fig. 2B). The resulting plasmid DNAs were transcribed with bacteriophage T7 RNA polymerase and were subsequently translated in a reticulocyte system. As shown in Figure
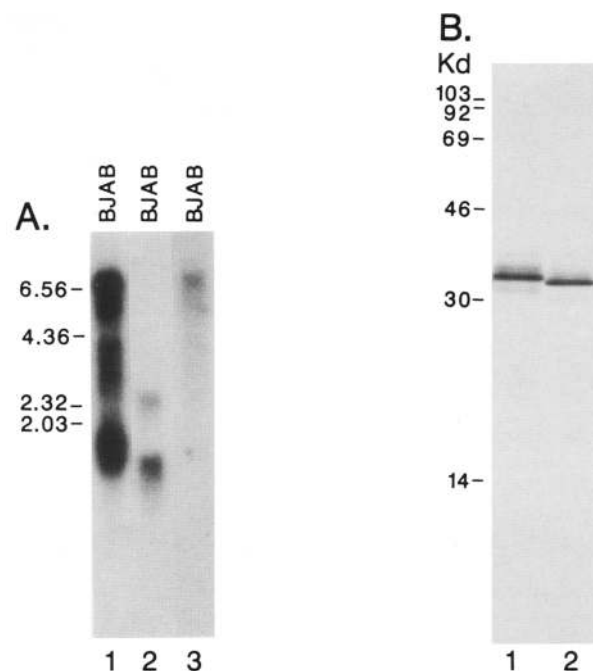
Figure 2. Northern analysis of poly(A)+-containing mRNAs from the human B-cell line BJAB. The filter was hybridized in lane *1* with the 1.1-kb cDNA from plasmid 3-1, in lane *2* with the 0.3-kb-long *Sma*I DNA fragment from plasmid pass-3 (position 1546 in Fig. 1B to the *Sma*I site in the pGEM4 polylinker at the 3' end of the cDNA), and in lane *3* with the 0.12-kb-long *Sac*I DNA fragment from plasmid 3-1 (position 1613 to the *Sac*I site in the polylinker). The blot was hybridized at 42°C in presence of formamide and washed under stringent conditions (0.2× SSC/65°C). Numbers on the left indicate the lengths (in kb) of comigrated RNA standards. (*B*) Analysis by SDS-PAGE (20% PAGE) of polypeptides translated from the 3' regions of the cDNAs in plasmids pass-3 (lane *1*) or 3-1 (lane *2*). (Lane *1*) A *Pst*I–*Eco*RI DNA fragment from vector pBS-ATG–oct-2 (position 1137–1713 in Fig. 1A; see Fig. 5A for description) was replaced by a *Pst*I–*Eco*RI fragment (position 1137 in Fig. 1A to position 1845 in Fig. 1B) from clone pass-3. DNA sequencing confirmed that the inserted pass-3 DNA fragment was in the same frame as in clone 3-1; (lane *2*) vector pBS-ATG–oct-2 (see Fig. 5A) was used for in vitro transcription/translation.

2B, the resulting polypeptides migrated with the mobilities of the anticipated molecular weights 34 and 32.4 kD. The polypeptide from the pass-3 cDNA was 1.6 kD larger than that from the 3-1 cDNA. Both polypeptides specifically bound a probe containing the octanucleotide sequence, producing a readily detectable DNA–protein complex in the gel mobility assay (data not shown). This suggests that the oct-2 gene is expressed as a family of polypeptides in B cells.

Evidence that the oct-2 gene encodes NF-A2

The NF-A2 factor was purified from nuclear extracts of BJAB cells by conventional chromatography followed by multiple passages over an affinity column containing immobilized oligomers of the octanucleotide sequence (Fig. 3A, lane 2). The purified NF-A2 consisted of three bands, as resolved by gel electrophoresis: a major band and two minor bands with deduced molecular masses of 61, 58, and 63 kD, respectively. A cDNA (pass-3) for the *oct-2* gene was inserted into the polylinker of the pGEM (Promega) expression vector. Translation of RNA transcribed from the SP6 promoter–pass-3 cDNA construct yielded a major polypeptide of 61 kD, which comigrated with the prominent polypeptide from the purified sample of NF-A2 (Fig. 3A, lane 3).

The mobility of a DNA–protein complex in the gel assay is determined primarily by the molecular weight of the protein. Complexes were generated with the affinity-purified NF-A2 and the products of translation in vitro of RNA from the oct-2 cDNA. These complexes comigrated during electrophoresis in a native gel (Fig. 3B, lanes 7 and 8), again suggesting that the oct-2 cDNA encodes the major form of the NF-A2 factor.

The affinity-purified NF-A2 protein and the polypeptide translated in vitro from the oct-2 cDNA were also compared by partial tryptic digestion. Samples from different digestion times of NF-A2 were resolved by denaturing gel electrophoresis and detected by staining with silver (Fig. 4, lanes 1–6). The mobility of these partial fragments was compared with those observed after a parallel analysis of [35S]methionine-labeled polypeptide from transcription/translation of the pass-3 cDNA in vitro (Fig. 4, lanes 7–12). The two samples generated a similar set of digestion fragments, again suggesting that NF-A2 is encoded by the oct-2 cDNA.

The sequence-specific binding domain of Oct-2

Protein sequence comparisons suggested that the DNA-binding domain of Oct-2 was specified by a domain (positions 955–1134) that was related distantly to both the helix-turn-helix structure of bacterial repressors and the homeo box proteins. To test this analogy directly, a fragment of the cDNA encompassing this region (655–1713) was inserted into the expression vector pBS-ATG so that RNA could be transcribed from the truncated templates by bacteriophage T7 RNA polymerase, as indicated in Figure 5A. The polypeptides translated in vitro from these RNAs were tested for specific DNA binding by the addition of the total translation mix to the DNA–protein gel assay. Polypeptides produced from RNAs terminating at positions 1713 (*Kpn*I site in the pBS-ATG polylinker), 1446 (*Stu*I), and 1137 (*Pst*I) specifically bound the octanucleotide-containing probe (Fig. 5B, lanes 1–3), whereas the polypeptide translated from RNA terminating at the 948 (*Eag*I) site did not bind specifically. The
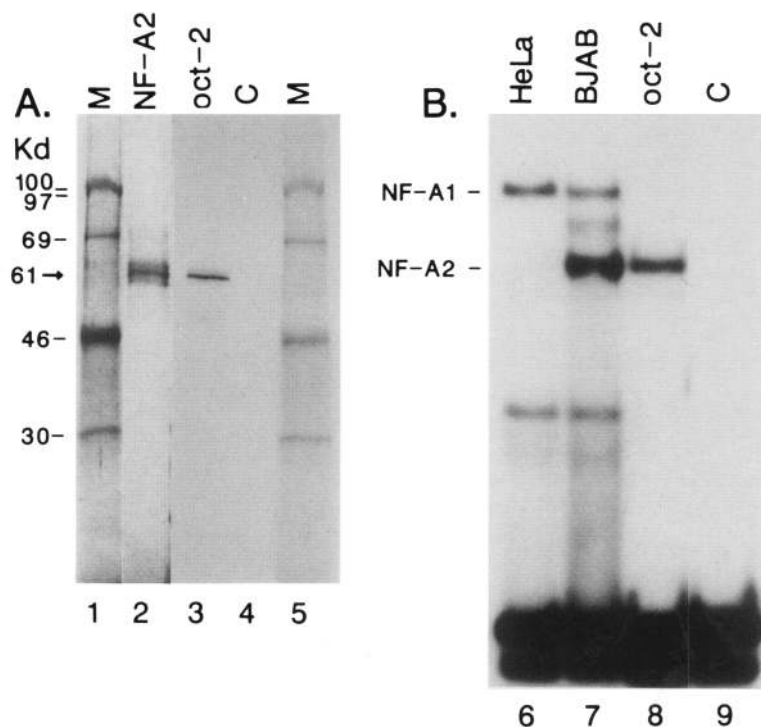
Figure 3. (*A*) Analysis by SDS-PAGE of affinity-purified NF-A2 from the B-cell line BJAB (lane 2) and of protein translated in vitro from *oct-2* gene-derived cDNA (lane 3). A pGEM–pass-3 vector was constructed by insertion of the *Eco*RI-linker-tailed pass-3 cDNA insert (positions 92–1845 in Fig. 1A,B, respectively) into the *Eco*RI site of the polylinker of the pGEM4 vector (Promega). (Lane 3) In vitro-translated [35S]methionine-labeled Oct-2 protein was generated from RNAs transcribed from the SP6 promoter in the pGEM–pass-3 vector. The NF-A2 proteins were silver stained (lane 2), and the [35S]methionine-labeled Oct-2 protein was detected by autoradiography. (*B*) Gel mobility shift assay of nuclear extracts of HeLa cells (lane 6), BJAB cells (lane 7), and reticulocyte lysates containing translation products from the pass-3 cDNA segment, described in *A* (lane 8), and unreacted reticulocyte (lane 9). The standard probe containing the octanucleotide sequence (ATGCAAAT) was used in lanes 6–9.
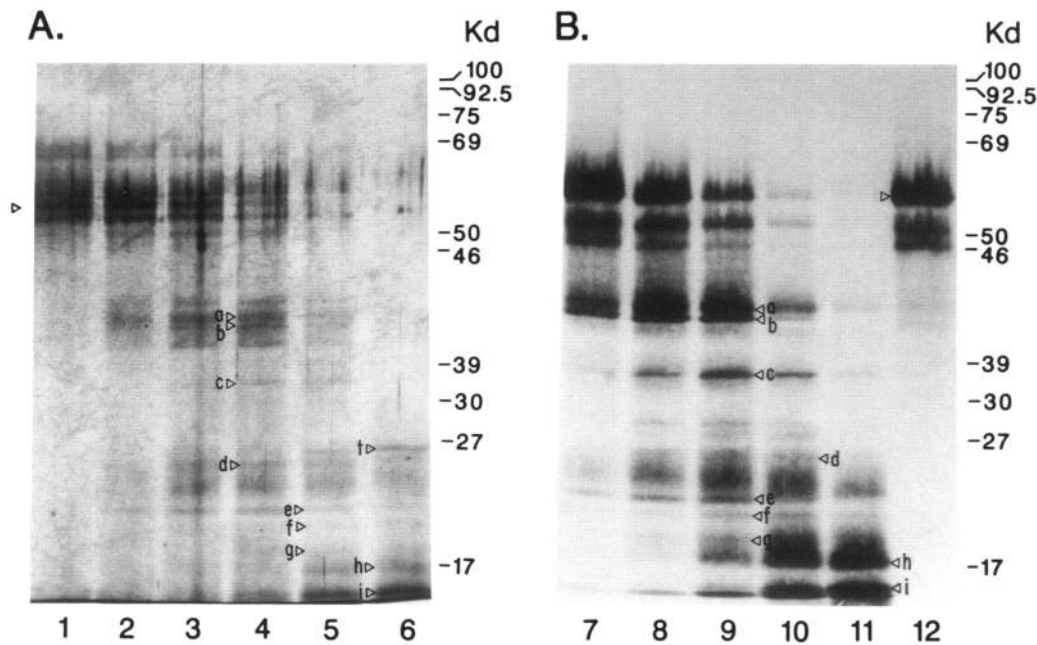
Figure 4. Analysis by SDS-PAGE of tryptic digestions of affinity purified from BJAB cells NF = A2 protein (*A*) and of [³⁵S]methionine-labeled Oct-2 protein translated in vitro from pass-3 cDNA (*B*). (*A*) The proteins were silver stained; (*B*) the [³⁵S]methionine-labeled proteins were detected by autoradiography. (Lane *1*) NF-A2 protein; (lanes *2–6*) NF-A2 digested with 3, 6, 12, 24, or 48 ng of trypsin, respectively; (lanes *7–11*) affinity-purified Oct-2 protein synthesized with [³⁵S]methionine, as described in Fig. 4A, reacted with 3, 6, 12, 24, and 48 ng of trypsin, respectively; (lane *12*) Oct-2 protein. Tryptic peptides a–i migrate at the same rate in *A* and *B*. (t) A band due to trypsin.

region containing the helix-turn-helix portion of Oct-2 is deleted in the latter protein. Because the truncated polypeptide encoded by RNA from the latter template was translated efficiently in the reticulocyte reaction (data not shown), this suggests that the specific binding of the Oct-2 protein requires the helix-turn-helix structure.

Two distinct but similarly migrating protein–DNA complexes were detected in the sample generated by translation of RNA from the *Stu*I-cleaved template (Fig. 5B, lane 2). The faint slower-migrating complex comigrated with the complex generated with templates cleaved by *Kpn*I (lane 1). The presence of the two complexes in the *Stu*I sample is due to a partial digestion of the plasmid DNA. The slower-migrating complex is probably produced by protein terminated at the stop codon TAA located at position 1468. The faster-migrating complex probably results from molecules terminated at the *Stu*I site. This interpretation was supported by the resolution of two ³⁵S-labeled polypeptides during gel electrophoresis of the *Stu*I sample and confirms the position of the termination codon of Oct-2.

Many sequence-specific binding proteins have an oligomeric structure. For example, bacterial repressor proteins typically bind sites with twofold rotational symmetry by forming a similarly symmetric dimer (Ptashne 1986). It should be noted that the binding site sequence of the Oct-2 protein is not symmetric, but oligomeric proteins could bind to nonsymmetric sites. Other examples of oligomerization of sequence-specific

binding proteins are the GCN4 protein of yeast (Hope and Struhl 1987) and the C/EBP protein of mammals. In the latter case, an α-helical region with an amphipathic character reflected in the spacing of four leucine residues by exactly seven residues is thought to be responsible for dimer formation (Landschultz et al. 1988a,b). A convenient assay for detection of dimerization of sequence-specific binding proteins is to cotranslate RNAs encoding two different size forms of the protein and test whether protein–DNA complexes with novel mobilities are generated (Hope and Struhl 1987). If only monomers bind to the probe, the sample containing the cotranslated polypeptides will generate only the complexes detected when either RNA is assayed singularly. This was the case with combinations of different-length RNAs transcribed from the oct-2 cDNA segment. Specifically, cotranslation of RNAs from templates cleaved at *Stu*I (1446) and *Pst*I (1137) did not generate novel bands in the gel mobility assay (Fig. 5B, lane 6). Similarly, cotranslation of two RNAs transcribed from the pass-3 cDNA and the 3-1 cDNA (pBS-ATG–oct-2) template, respectively, and encoding different size proteins (58 and 32 kD) containing both the DNA-binding region and the potential leucine zipper region failed to generate novel bands in the gel mobility shift assay (data not shown). Thus, on the basis of this negative evidence, we suggest that a single molecule of the Oct-2 protein is present in the resolved DNA–protein complexes and that it does not require dimerization for binding to DNA.
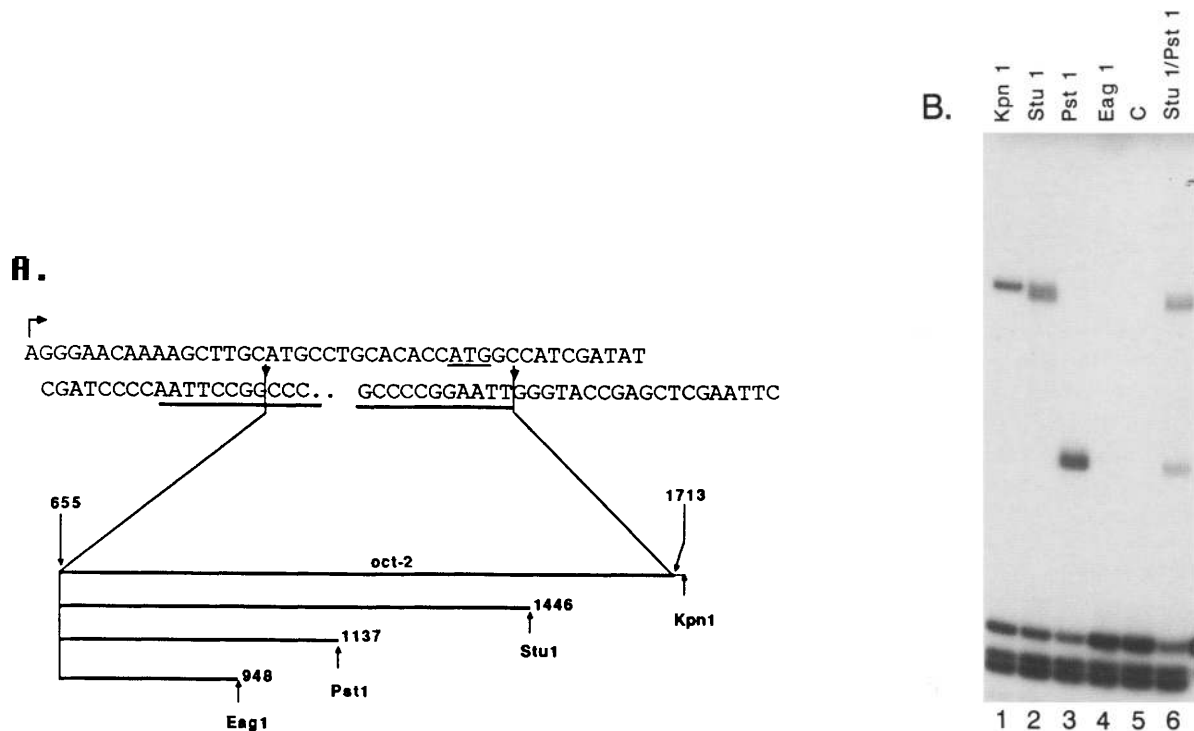
Figure 5. (A) Schematic of expression plasmid pBS-ATG−oct-2. A 27-bp oligonucleotide containing an ATG codon surrounded by the appropriate bases for efficient initiation (Kozak 1986) was ligated into the PstI site of the polylinker of Bluescript pBS+/− plasmid (Stratagene, Inc.). The 1.1-kb-long cDNA in plasmid 3-1, positions 655−1713 in Fig. 1A, was ligated into the SmaI site of this vector. The two lines of nucleotide sequence are contiguous. The arrow indicates the site of initiation of transcription. Initiation of translation is anticipated at the underlined ATG codon. Numbers refer to nucleotide positions described in Fig. 1A. (B) Gel mobility shift assays of Oct-2 proteins synthesized in vitro from truncated templates. The sites where the template was cleaved by restriction enzymes to produce truncated RNA (KpnI, StuI, PstI, and EagI) are indicated. (Lanes 1−6) The translation products were assayed with a wild-type octamer DNA probe (ATGCAAAT).

Discussion

The proteins encoded by the oct-2 cDNAs undoubtedly correspond to the B-cell-specific nuclear factor NF-A2 (OTF-2), identified previously by octanucleotide DNA−protein binding assays (Singh et al. 1986; Staudt et al. 1986). This protein will be referred to as the Oct-2 protein, or factor. Expression of oct-2 in B cells is probably responsible for specific transcription of the immunoglobulin heavy- and light-chain genes. At the moment, oct-2 expression is thought to be restricted to B cells, but studies of other cell types have been limited.

The sequence of the Oct-2 protein has several interesting features. In common with many transcription factors, it contains multiple regions high in glutamine and one particular region high in acidic residues. Such acidic regions in other factors have been shown to be important in activation of transcription (Gill and Ptashne 1987; Hope et al. 1988). Studies of transactivation of immunoglobulin promoters by mutant genes constructed from the oct-2 cDNA will help to determine the functions of these regions. Surprisingly, the Oct-2 protein also contains a potential leucine zipper motif, where four leucine residues are separated by exactly seven amino acids. This distribution of leucine residues permits the theoretical formation of an amphipathic α-helix, which can pair by interaction of the aligned hydrophobic leucine residues (Landschultz et al. 1988a). Consistent with this suggestion, no helix-disrupting proline residue is present in Oct-2 in the 22-amino-acid tract defined by the four leucines. However, unlike the first example of a leucine zipper protein, C/EBP, the potential α-helical region in Oct-2 does not possess a high density of paired charged residues that could stabilize the structure. Also, unlike the C/EBP protein, which binds DNA specifically as a homodimer, probably by pairing through the leucine zipper, the Oct-2 protein appears to bind DNA specifically as a monomer. It is interesting to speculate that the leucine zipper region of Oct-2 might be important for interactions with other proteins, as there is no obvious reason to restrict the binding of such a structure to self-recognition.

The domain of Oct-2 involved in specific DNA binding has homology to the helix-turn-helix structure first identified in repressors of bacteriophage and to the homeo box sequence first identified in developmental regulatory genes of Drosophila. The 60-residue homeo box domain is highly conserved among 16 examples in different Drosophila genes (Gehring 1987). This level of

conservation extends to homeo box sequences found in vertebrates and worms. Among this total family, 9 of the 60 residues are invariant. The Oct-2 protein only contains six of these nine residues, and four of these six sites are clustered in the subregion of the homeo box thought to be related to the helix-turn-helix structure (see Fig. 6). A number of yeast regulatory proteins, MATa1, MATα2, pho2, and a *Caenorhabditis elegans* differentiation factor, *mec-3*, also have homology through this subregion of the homeo box but do not conserve the other invariant residues of the homeo box. Thus, we suggest that Oct-2, these yeast proteins, and homeo box proteins are primarily related through a common structure involved in DNA binding and that the more extended conservation among the original set of homeo box proteins reflects a common biological function of these factors that the Oct-2 factor probably does not possess.

As we have described previously, a constitutive factor, NF-A1, is present in all cell types and recognizes the octanucleotide sequence with a specificity indistinguishable from that of Oct-2 (Staudt et al. 1986). In fact, single point mutations throughout the octanucleotide sequence affected the binding of these two factors equivalently. This constitutive factor probably regulates transcription of the H2B histone gene and the snRNA genes in all cell types. In an accompanying paper, Sturm et al. (1988, this issue) report the sequence of a cDNA segment and the evidence that it encodes the constitutive factor (NF-A1, OTF-1, or OBP100). This protein encoded by this cDNA is referred to as the Oct-1 protein.

Homology between the Oct-2 and Oct-1 proteins is illustrated in Figure 7. This alignment suggests that the two proteins share the three high-glutamine regions and the acidic cluster. However, Oct-1 does not contain sequences homologous to the leucine zipper region of Oct-2. Most strikingly, the two proteins have 74 of 75 identical amino acids between positions 649 and 873 upstream of the homeo box-related domain and 53 of 60 amino acids in the homeo box-related region. The sequence between these two highly conserved regions is significantly less conserved, 50% (13 of 26 amino
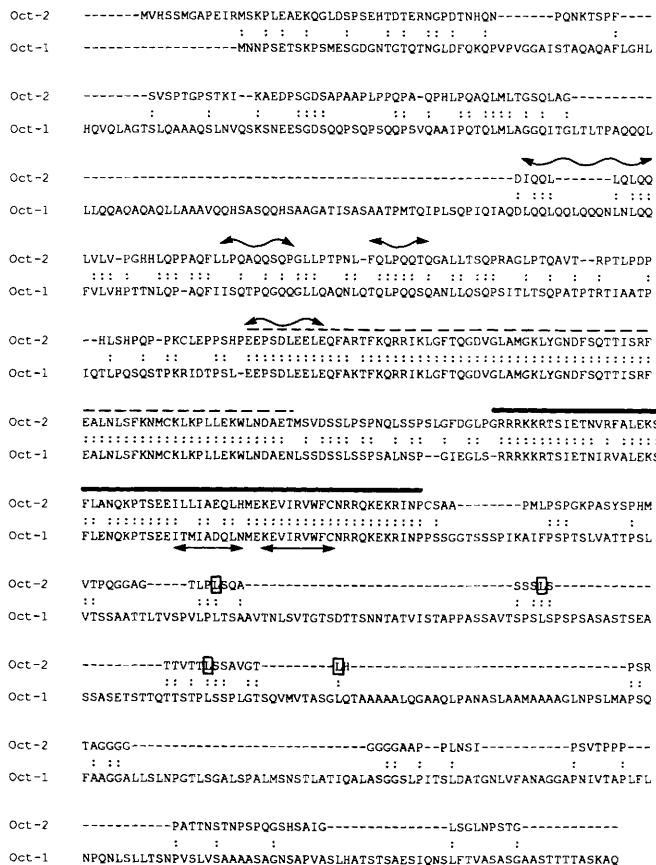
acids). Furthermore, the two proteins show, at best, only short patches of potential homology beyond the boundaries of the highly conserved regions. This lack of sequence conservation in flanking regions suggests that these two genes are related very distantly. Thus, the extent of similarity between Oct-1 and Oct-2 in the two conserved regions must reflect an important common function, or functions. One of these functions is specific binding to the octanucleotide sequence. In view of the high homology of the region containing the helix-turn-helix structure, it is not surprising that the binding specificities of these two proteins are difficult to resolve. It is unlikely that all the sequence conservation in the homeo box-related region of these two proteins simply reflects constraints due to specific sequence recognition. More likely, these two proteins probably share another common function in the amino-terminal portion of the homeo box-related domain. In contrast, the proteins in the highly related homeo box family probably share a common function specified by this region that is not shared by Oct-1 and Oct-2 proteins. The nature of these functions can only be subjects of speculation. However, the most likely type of function is interaction with other proteins important in gene regulation.

The remarkably conserved 75-residue region that is amino-terminal to the homeo box-type domain must specify another function common to Oct-1 and Oct-2. This function probably does not involve sequence-specific DNA binding, as deletion of the homeo box-related domain eliminates binding by Oct-2, and other families of homeo box-related domains do not contain a similarly conserved region. At present, there are no known common activities of Oct-1 and Oct-2 other than DNA binding and activation of transcription (LeBowitz et al. 1988). In fact, the interesting issue is the differences in the biology of these two regulatory proteins. The Oct-1 protein is thought to interact specifically with other DNA-binding transcription proteins to promote transcription of genes in all cell types. Furthermore, Oct-1 activity is probably regulated in a cell-cycle-specific fashion (Fletcher et al. 1987; LaBella et al. 1988). In contrast, Oct-2 is thought to be uniquely active in immuno-
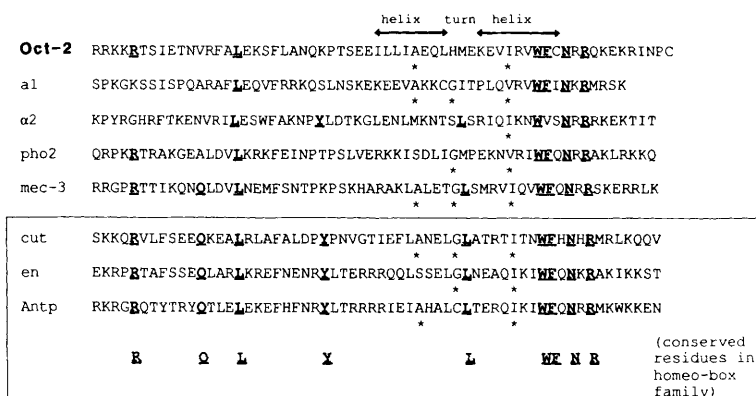
Figure 6. Amino acid sequence alignment of the DNA-binding domain of Oct-2 factor with homeo boxes from *Antp* (Schneuwly et al. 1986), *cut* (Blochlinger et al. 1988), and *engrailed* (*en*) (Poole et al. 1985) proteins (boxed-in amino acid sequences), and with homeo box-related amino acid sequences from the *Saccharomyces cerevisiae* proteins MATa1 (Miller 1984), MATα2 (Astell et al. 1981); *pho2* (Sengstag and Hinnen 1987; Bürglin 1988), and *C. elegans* protein *mec-3* (Way and Chalfie 1988). The nine invariant residues in canonical homeo box sequences *Antp*, *cut*, and *en* are listed below the boxed-in amino acid sequences and shown in boldface type if present in the amino acid sequences. Stars indicate the hydrophobic amino acids that are critical for the protein to maintain the helix-turn-helix structure (Pabo and Sauer 1984). Solid arrows delimit the helix-turn-helix domain.

```
Oct-2    -------MVHSSMGAPEIRMSKPLEAEKQGLDSPSEHTDTERNGPDTNHQN-------PQNKTSPF----
              : : : :   :   : : ::  ::  :             :
Oct-1    ------------------MNNPSETSKPSMESGDGNTGTQTNGLDFQKQPVPVGGAISTAQAQAFLGHL


Oct-2    --------SVSPTGPSTKI--KAEDPSGDSAPAAPLPPQPA-QPHLPQAQLMLTGSQLAG---------
              :       :    :  ::::       :: :  :: :::: : : :
Oct-1    HQVQLAGTSLQAAAQSLNVQSKSNEESGDSQQPSQPSQQPSVQAAIPQTQLMLAGGQITGLTLTPAQQQL

                                                        ⌢⌣⌢
Oct-2    ----------------------------------------------------DIQQL-------LQLQQ
                                                             : :::     :  : :::
Oct-1    LLQQAQAQAQLLAAAVQQHSASQQHSAAGATISASAATPMTQIPLSQPIQIAQDLQQLQQLQQQNLNLQQ

          ⌢⌣⌢             ⌢⌣
Oct-2    LVLV-PGHHLQPPAQFLLPQAQQSQPGLLPTPNL-FQLPQQTQGALLTSQPRAGLPTQAVT--RPTLPDP
         :::: :  ::: ::: : : : :::   ::  ::::::  :  :: :::    :  :  :     :
Oct-1    FVLVHPTTNLQP-AQFIISQTPQGQQGLLQAQNLQTQLPQQSQANLLQSQPSITLTSQPATPTRTIAATP

           ⌢⌣⌢
Oct-2    --HLSHPQP-PKCLEPPSHPEEPSDLEELEQFARTFKQRRIKLGFTQGDVGLAMGKLYGNDFSQTTISRF
           : : :  :  ::  ::::::::::::::  ::::::::::::::::::::::::::::::::::::::::
Oct-1    IQTLPQSQSTPKRIDTPSL-EEPSDLEELEQFAKTFKQRRIKLGFTQGDVGLAMGKLYGNDFSQTTISRF

         ─ ─ ─ ─ ─ ─ ─ ─ ─ ─ ─
Oct-2    EALNLSFKNMCKLKPLLEKWLNDAETMSVDSSLPSPNQLSSPSLGFDGLPGRRRKKRTSIETNVRFALEKS
         :::::::::::::::::::::: : :::: :: :  :  : :: ::::::::::::: : :::::
Oct-1    EALNLSFKNMCKLKPLLEKWLNDAENLSSDDSSLSSPSALNSP--GIEGLS-RRRKKRTSIETNIRVALEKS

         ────────────────
Oct-2    FLANQKPTSEEILLIAEQLHMEKEVIRVWFCNRRQKEKRINPCSAA--------PMLPSPGKPASYSPHM
         :: ::::::::  :: :: ::::::::::::::::::::: :           :
Oct-1    FLENQKPTSEEITMIADQLNMEKEVIRVWFCNRRQKEKRINPPSSGGTSSSPIKAIFPSPTSLVATTPSL
                                                 ⟵─────⟶  ⟵──────⟶

Oct-2    VTPQGGAG-----TLP[L]SQA---------------------------------SSS[L]S------------
         ::          ::: :                                     :  :::
Oct-1    VTSSAATTLTVSPVLPLTSAAVTNLSVTGTSDTTSNNTATVISTAPPASSAVTSPSLSPSPSASASTSEA


Oct-2    -----------TTVTT[L]SSAVGT---------[L]H-----------------------------PSR
                   :: :::: ::            :
Oct-1    SSASETSTTQTTSTPLSSPLGTSQVMVTASGLQTAAAAALQGAAQLPANASLAAMAAAAGLNPSLMAPSQ


Oct-2    TAGGGG--------------------------GGGGAAP--PLNSI----------PSVTPPP---
         : :::                         ::  :            :       :
Oct-1    FAAGGALLSLNPGTLSGALSPALMSNSTLATIQALASGGSLPITSLDATGNLVFANAGGAPNIVTAPLFL


Oct-2    -----------PATTNSTNPSPQGSHSAIG---------------LSGLNPSTG-----------
             :   :    :        :            :       :
Oct-1    NPQNLSLLTSNPVSLVSAAAASAGNSAPVASLHATSTSAESIQNSLFTVASASGAASTTTTASKAQ
```

Figure 7. Best-fit alignment of Oct-2 and Oct-1 (Sturm et al. 1988) predicted amino acid sequences, according to the IALIGN program from the University of Wisconsin Computer Group. Wavy lines delimit the glutamine (Q)-rich and acidic (E)- or (D)-rich regions. The POU box is depicted by a dotted line, the solid line marks the region related to the homeo box. Solid arrows delimit the helix-turn-helix domain. Boxed leucine (L) residues are spaced exactly by seven amino acids.

globulin gene regulation. Obviously, these two proteins share, as yet, one or more unrecognized common function that is specified by the highly conserved domain.

As discussed in Herr et al. (1988, this issue), the two highly conserved regions between Oct-1 and Oct-2 also have a high degree of homology—64% and 50%, or 60% and 46%, respectively—to comparable-length regions of the rat pituitary regulatory protein Pit-1 (Ingraham et al. 1988; Nelson et al. 1988) and the protein encoded by the unc-86 gene of Caenorhabditis elegans (Finney et al. 1988), respectively. This degree of conservation among genes in organisms as diverse as worms and mammals compares well with that observed in homeo box-containing genes. In fact, in the homeo box-related region, these four proteins are more related to one another than to other homeo box proteins. Again, this suggests a common function distinct from that of the other homeo box proteins. Similarly, for the more conserved amino-terminal region, these four proteins must share a

common function. [This highly conserved amino-terminal 75-amino-acid region shared by the Pit-1, Oct-1, Oct-2, and Unc-86 factors has been designated as a POU box (see Herr et al. 1988).] The Pit-1 factor is expressed only in pituitary cells and regulates the expression of the prolactin and growth hormone genes. The unc-86 gene regulates differentiation of cells in many lineages of the nematode. This suggests that the common set of functions shared by these proteins may involve some aspect of control of gene expression in differentiated cellular states.

The oct-2 gene must specify at least two different-length mRNAs, as two different cDNA segments were isolated and sequenced. These segments are identical through a majority of their sequences but abruptly diverge in sequence near their 3' ends. The two proteins encoded by these cDNAs differ, with one containing an additional 12 amino acids at its carboxyl terminus. The potential significance of these additional sequences is unclear. These two cDNAs almost certainly differ by alternative splicing patterns of RNA transcribed from the oct-2 gene. Furthermore, it is likely that the oct-2 gene encodes a more diverse set of mRNAs than those partially defined by these two cDNAs. Six different-length mRNAs are produced at significant levels in mature B cells. The relative amounts of these mRNAs vary between pre-B, B, and plasma cell lines (Staudt et al. 1988). This population could reflect variations in sites of initiation of transcription and of polyadenylation, as well as further differences in splicing patterns.

Perhaps the most unanticipated aspect of the sequence of the oct-2 cDNA was the presence of an overlapping second ORF, some 278 amino acids in length. Such overlapping reading frames have been described previously in viral genomes where compaction of sequences is critical for encapsidation (Sanger et al. 1978). However, few have been described in cDNAs of vertebrates and none have been shown to encode a functional protein. The probability in a random sequence of finding an ORF of 278 triplets in a sequence of the length of the oct-2 cDNA is $<10^{-4}$. The sequence of a putative protein encoded by this reading frame is quite rich in both charged amino acids and proline residues. In this superficial fashion, it is similar to the trans-activator proteins encoded by the E1A region of adenovirus. It is possible that the oct-2 gene represents a limited sequence complexity module designed by nature for control of B-cell development. The potential overlapping reading structure of the oct-2 gene may be explained by the advantages of compacting sequences for multiple proteins into one transcriptional unit that is regulated by other genes to be expressed only in B cells.

Materials and methods

DNA sequencing

DNA sequencing was performed on double-stranded plasmid DNA templates according to the Sanger dideoxynucleotide protocol, as modified by U.S. Biochemical for use with bacteriophage T7 DNA polymerase (Sequenase). The entire sequence was confirmed by sequencing the opposite strand and, in the

GC-rich regions, by sequencing according to Maxam and Gilbert (1980).

Plasmid constructions

cDNAs were subcloned from λgt11 to pGEM4 (Promega), and these plasmids were used for DNA sequence analysis and in vitro transcription. Plasmid pBS-ATG was kindly provided by H. Singh and K. LeClair and was constructed by ligating a 27-bp-long oligonucleotide TGCACACCATGGCCATCGA-TATCGATC into the *PstI* site of pBS – / + Bluescript plasmid (Stratagene). The expression vector pBS-ATG–oct-2, depicted in Figure 5A, was designed for transcription and translation in vitro and was constructed by cleaving pBS-ATG with *SmaI* and ligating the blunt-ended *EcoRI* 1.1-kb cDNA fragment from plasmid 3-1 (positions 655–1713 in Fig. 1A).

In vitro transcription/translation

In vitro transcription and translation reactions were performed as recommended by the manufacturer (Promega).

DNA-binding assay

The *EcoRI–HindIII* 50-bp fragment containing the wild-type octanucleotide sequence ATGCAAAT in the *BamHI* site of pUC18 polylinker was labeled with ^{32}P (50,000 cpm/ng), and 1 ng of DNA probe was incubated with 1 μl of the reacted/unreacted reticulocyte lysate. The binding reactions were incubated at room temperature for 30 min and contained 10 mM Tris-HCl (pH 7.5), 50 mM NaCl, 1 mM dithiothreitol (DTT), 1 mM EDTA (pH 8), 5% glycerol, and 25 μg/ml sonicated denatured calf thymus DNA in 2.5 μg/ml sonicated native calf thymus DNA as nonspecific competitors. The complexes were resolved by electrophoresis in 4% polyacrylamide gel (acrylamide/*bis*-acrylamide weight ratio of 29 : 1), containing as buffer 25 mM Tris-HCl (pH 8.5), 190 mM glycine, and 1 mM EDTA buffer, as described previously (Singh et al. 1986).

Purification of NF-A2

NF-A2 was purified to ≥90% homogeneity from nuclear extracts derived from the human Burkitt's lymphoma cell line, BJAB. Purification was accomplished by sequential fractionation on Zetachrom QAE discs (Cuno, Inc.), heparin Sepharose (Pharmacia), single-stranded (ss) DNA cellulose (Pharmacia), and a DNA affinity column that contained an immobilized double-stranded (ds) segment containing the octanucleotide sequence (LeBowitz et al. 1988). Details of the purification will be published elsewhere (J.H. LeBowitz, unpubl.). In vitro-translated, [^{35}S]methionine-labeled Oct-2 protein was purified by chromatography on dsDNA cellulose followed by affinity chromatography on the octanucleotide DNA affinity column.

Tryptic digestion of NF-A2 and Oct-2 protein

Tryptic digests were performed at room temperature in a buffer consisting of 20 mM HEPES–KOH (pH 7.9), 20% glycerol, 0.5 M KCl, 0.2 mM EDTA, and 0.5 mM DTT. Aliquots of purified NF-A2 (~250 ng) or affinity-purified Oct-2 protein (90,000 cpm) were incubated with varying amounts of trypsin (affinity-purified trypsin was gift from Dan Doering). After 60 min, reactions were terminated by the addition of 2.5 volumes of SDS-PAGE sample buffer and were boiled for 5 min. The tryptic digests were resolved on 10% polyacrylamide gels. Tryptic fragments of NF-A2 were visualized by silver staining. Tryptic

fragments of [^{35}S]methionine-labeled Oct-2 protein were visualized by autoradiography after treatment of the gel with En^{3}hance (Dupont).

Note added in review

While this paper was in review, the sequence of the region of the oct-2 cDNA containing the homeo box-related structure was reported (Ko et al. 1988).

Note

Sequence data described in this paper have been submitted to the EMBL/GenBank Data Libraries under accession number Y00227.

Acknowledgments

References

Astell, C.R., L. Ashlstrom-Jonasson, M. Smith, K. Tatchell, K. Nasmyth, and D.B. Hall. 1981. The sequence of the DNAs coding for the mating type loci of *Saccharomyces cerevisiae*. *Cell* **27**: 15–23.

Banerji, J., L. Olson, and W. Schaffner. 1983. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell* **33**: 729–739.

Bergman, Y., D. Rice, R. Grosschedl, and D. Baltimore. 1984. Two regulatory elements for immunoglobulin κ light chain gene expression. *Proc. Natl. Acad. Sci.* **81**: 7041–7045.

Blochlinger, K., R. Bodmer, J. Jack, L. Yeh Jan, and Y.N. Jan. 1988. Primary structure and expression of a product from *cut*, a locus involved in specifying sensory organ identity in *Drosophila*. *Nature* **333**: 629–635.

Bürglin, T.R. 1988. The yeast regulatory gene *pho2* encodes a homeo box. *Cell* **53**: 339–340.

Carbon, P., S. Murgo, J.P. Ebel, A. Krol, G. Tebb, and I.W. Mattaj. 1987. A common octamer motif binding protein is involved in the transcription of U6 snRNA by RNA polymerase III and U2 snRNA by RNA polymerase II. *Cell* **51**: 71–79.

Eaton, S. and K. Calame. 1987. Multiple DNA sequence elements are necessary for the function of an immunoglobulin heavy chain promoter. *Proc. Natl. Acad. Sci.* **84**: 7634–7638.

Falkner, F.G. and H.Z. Zachau. 1984. Correct transcription of an immunoglobulin κ gene requires an upstream fragment

containing conserved sequence elements. *Nature* **310**: 71–74.

Finney, M., G. Ruvkun, and H.R. Horvitz. 1988. The *C. elegans* cell-lineage and differentiation gene *unc-86* encodes a protein with a homeo domain and extended sequence similarity to mammalian transcription factors *Cell* (in press).

Fletcher, C., N. Heintz, and R.G. Roeder. 1987. Purification and characterization of OTF-1, a transcription factor regulating cell cycle expression of a human histone H2B gene. *Cell* **51**: 773–781.

Gehring, W.J. 1987. Homeo boxes in the study of development. *Science* **236**: 1245–1252.

Gerster, T., P. Matthias, M. Thali, J. Jiricny, and W. Schaffner. 1987. Cell type-specificity elements of the immunoglobulin heavy chain gene enhancer. *EMBO J.* **6**: 1323–1330.

Gill, F. and M. Ptashne. 1987. Mutants of GAL4 protein altered in an activation function. *Cell* **51**: 121–126.

Gillies, S.D., S.L. Morrison, V.T. Oi, and S. Tonegawa. 1983. A tissue-specific transcriptional enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. *Cell* **33**: 717–728.

Grosschedl, R. and D. Baltimore. 1985. Cell type specificity of immunoglobulin gene expression is regulated by at least three DNA sequence elements. *Cell* **41**: 885–897.

Herr, W., R. Sturm, R.G. Clerc, L.M. Corcoran, D. Baltimore, P.A. Sharp, H.A. Ingraham, M.G. Rosenfeld, M. Finney, G. Ruvkun, and H.R. Horvitz. 1988. The POU domain a large conserved region in the mammalian *pit-1*, *oct-1*, *oct-2*, and *Caenhabditis elegans unc-86* gene products. *Genes Dev.* **2**: 1513–1516.

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.

Hope, I.A., S. Mahadevan, and K. Struhl. 1988. Structural and functional characterization of the short acidic transcriptional activation region of yeast GCN4 protein. *Nature* **333**: 635–640.

Ingraham, H.A., R. Chen, H.J. Mangalain, H.P. Elsholtz, S.E. Flynn, C.R. Lin, D.M. Simmons, L. Swanson, and M.G. Rosenfeld. 1988. A tissue-specific transcription factor containing a homeo domain specifies a pituitary phenotype. *Cell* (in press).

Ingraham, H.A. and M.G. Rosenfeld. 1988. A human protein specific for the immunoglobulin octamer DNA motif contains a functional homeo box domain. *Cell* (in press).

Ko, H.S., P. Fast, W. McBride, and L.M. Staudt. 1988. *Cell* **55**: 135–144.

Kozak, M. 1986. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* **44**: 283–292.

Krol, A., P. Carbon, J.P. Ebel, and B. Appel. 1987. Xenopus tropicalis U6 snRNA genes transcribed by RNA polymerase III contain the promoter elements used by polymerase II dependent U snRNA genes. *Nucleic Acids Res.* **15**: 2463–2478.

LaBella, F., H.L. Sive, R.G. Roeder, and N. Heintz. 1988. Cell-cycle regulation of a human histone H2B gene is mediated by H2B subtype-specific consensus element. *Genes Dev.* **2**: 32–39.

Landschultz, W.H., P.F. Johnson, and S.L. McKnight. 1988a. The leucine zipper: A hypothetical structure common to a new class of DNA binding proteins. *Science* **240**: 1759–1764.

Landschultz, W.H., P.F. Johnson, E.Y. Adashi, B.J. Graves, and S.L. McKnight. 1988b. Isolation of a recombinant copy of the gene encoding C/EBP. *Genes Dev.* **2**: 786–800.

LeBowitz, J.H., T. Kobayashi, L. Staudt, D. Baltimore, and P.A. Sharp. 1988. Octamer-binding proteins from B or HeLa cells

stimulate transcription of the immunoglobulin heavy chain promoter in vitro. *Genes Dev.* **2**: 1227–1237.

Mason, J.O., G.T. Williams, and M.S. Neuberger. 1985. Transcription cell type specificity confused by an immunoglobulin N_H gene precursor that includes a functional consensus sequence. *Cell* **41**: 479–487.

Maxam, A. and W. Gilbert. 1980. Sequencing end-labeled DNA with base-specific chemical cleavages. *Methods Enzymol.* **65**: 449–560.

Miller, A.M. 1984. The yeast MATa1 gene contains two introns. *EMBO J.* **3**: 1061–1065.

Nelson, C., V.R. Albert, H.P. Elsholtz, L.I.W. Lu, and M.G. Rosenfeld. 1988. Activation of cell-specific expression of rat growth hormone and prolactin genes by a common transcription factor. *Science* **239**: 1400–1405.

O'Neill, E.A. and T.J. Kelly. 1988. Purification and characterization of nuclear factor III (origin recognition protein C), a sequence specific DNA binding protein required for efficient initiation of adenovirus DNA replication. *J. Biol. Chem.* **263**: 931–937.

O'Neill, E.A., C. Fletcher, C.R. Burrow, N. Heintz, R.G. Roeder, and T.J. Kelly. 1988. Transcription factor OTF-1 is functionally identical to the DNA replication factor NF-III. *Science* **241**: 1210–1213.

Pabo, C.O. and R.F. Sauer. 1984. Protein-DNA recognition. *Annu. Rev. Biochem.* **53**: 293–321.

Parslow, T.G., D.L. Blair, W.J. Murphy, and D.K. Granner. 1984. Structure of the 5' ends of immunoglobulin genes: A novel conserved sequence. *Proc. Natl. Acad. Sci.* **81**: 2650–2654.

Peterson, C.L. and K.L. Calame. 1987. Complex protein binding within the mouse immunoglobulin heavy-chain enhancer. *Mol. Cell. Biol.* **7**: 4194–4203.

Poole, S.J., L.M. Kauver, B. Drees, and T. Kornberg. 1985. The engrailed locus of *Drosophila*: Structural analysis of an embryonic transcript. *Cell* **40**: 37–43.

Ptashne, M. 1986. *A genetic switch*. Cell Press and Blackwell Scientific Publications, Cambridge, Massachusetts.

Sanger, F., A.R. Coulson, T. Friedmann, G.M. Air, B.G. Barrell, N.L. Brown, J.C. Fiddes, C.A. Hutchinson, P.M. Slocombe, and M. Smith. 1978. The nucleotide sequence of bacteriophage φX174. *J. Mol. Biol.* **125**: 225–246.

Scheidereit, C., A. Heguy, and R.G. Roeder. 1987. Identification and purification of a human lymphoid octamer-binding protein (OTF-2) that activates transcription of an immunoglobulin promoter in vitro. *Cell* **51**: 783–793.

Schneuwly, S., A. Kuroiwa, P. Baumgartner, and W.J. Gehring. 1986. Structural organization and sequence of the homeotic gene *Antennapedia* of *Drosophila melanogaster*. *EMBO J.* **5**: 733–739.

Scott, M.P. and A.J. Weiner. 1984. Structural relationships among genes that control development: Sequence homology between the Antennapedia, Ultrabithorax, and fushi tarazu loci in *Drosophila*. *Proc. Natl. Acad. Sci.* **81**: 4115–4119.

Sengstag, C. and A. Hinnen. 1987. The sequence of the *Saccharomyces cerevisiae* gene pho2 codes for a regulatory protein with unusual amino acid composition. *Nucleic Acids Res.* **15**: 233–246.

Singh, H., J.H. LeBowitz, A.S. Baldwin, and P.A. Sharp. 1988. Molecular cloning of an enhancer binding protein: Isolation by screening of an expression library with recognition site DNA. *Cell* **52**: 415–423.

Singh, H., R. Sen, D. Baltimore, and P.A. Sharp. 1986. A nuclear factor that binds to a conserved sequence motif in transcriptional control elements of immunoglobulin genes. *Nature* **319**: 154–158.

Staudt, L.M., R.G. Clerc, H. Singh, J.H. LeBowitz, P.A. Sharp, and D. Baltimore. 1988. Cloning of a cDNA encoding a B-

cell restricted octamer binding factor. *Science* **241:** 577–580.

Staudt, L.M., H. Singh, R. Sen, T. Wirth, P.A. Sharp, and D. Baltimore. 1986. A lymphoid-specific protein binding to the octamer motif of immunoglobulin genes. *Nature* **323:** 640–643.

Sturm, R., T. Baumruker, B.R. Franza, and W. Herr. 1987. A 100-kD HeLa cell octamer-binding protein (OBP100) interacts differently with two separate octamer-related sequences within the SV40 enhancer. *Genes Dev.* **1:** 1147–1160.

Sturm, R.A., G. Das, and W. Herr. 1988. The ubiquitous octamer-binding protein Oct-1 contains a POU domain with a homeo box subdomain. *Genes Dev.* **2:** 1582–1599.

Way, J.C. and M. Chalfie. 1988. mec-3, a homeo box-containing gene that specifies differentiation of the touch receptor neurons in *C. elegans. Cell* **54:** 5–16.

Wirth, T., L. Staudt, and D. Baltimore. 1987. An octamer oligonucleotide upstream of a TATA motif is sufficient for lymphoid-specific promoter activity. *Nature* **329:** 174–177.

The B-cell-specific Oct-2 protein contains POU box- and homeo box-type domains.

R G Clerc, L M Corcoran, J H LeBowitz, et al.

References	This article cites 47 articles, 17 of which can be accessed free at: http://genesdev.cshlp.org/content/2/12a/1570.full.html#ref-list-1
License	
Email Alerting Service	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**