Molecular characterization of *bsg25D*: a blastoderm-specific locus of *Drosophila melanogaster*

Paul D.Boyer[1]*, Paul A.Mahoney[1] and Judith A.Lengyel[1,2]+

[1]Molecular Biology Institute and [2]Biology Department, University of California, Los Angeles, CA 90024, USA

Received September 30, 1986; Revised and Accepted January 27, 1987

ABSTRACT

The blastoderm stage of Drosophila embryogenesis is a time of crucial transitions in RNA transcription, the cell cycle and segment determination. We have previously identified three loci encoding RNAs specific to this stage (Roark et al., Dev. Biol. 109, 476-488, 1985). We present here the complete nucleotide sequence of one of these loci, bsg25D, which encodes a 2.7 kb blastoderm-specific RNA. The primary structure of this RNA, and that of an overlapping 4.5 kb RNA, has been determined. The amino acid sequence of the predicted bsg25D protein has been compared to the NBRF protein database. Structural similarities between domains in the bsg25D, fos, and tropomyosin proteins, and their possible significance for early embryogenesis are discussed.

INTRODUCTION

Dramatic transitions occur at the blastoderm stage (1.5-3.5 hrs after fertilization) of Drosophila embryogenesis (reviewed in 1). Nuclei, which have been dividing synchronously at the highest rate known for eukaryotes, migrate from the interior of the embryo to its surface to form the syncytial blastoderm, become surrounded by membranes to generate the cellular blastoderm, and traverse the first true cell cycle. RNA transcription is activated to the highest embryonic level, per nucleus, during this time, and by the end of the blastoderm stage, cells have become determined as to their segmental fate in the ectoderm of the larva and adult.

One approach to understanding these events is the isolation and characterization of genomic DNAs encoding mRNAs specific to the blastoderm stage. Three blastoderm-specific genes (i.e., genes encoding RNAs which are 50-100 times more abundant in blastoderm embryos than at any other stage) have been identified by molecular screening techniques (2). This approach has identified two loci which encode proteins with putative "DNA-binding fingers" (3, reviewed in 4, Baldarelli et al., in preparation); these genes may be involved in the regulation of other genes at the blastoderm stage.

We present here the molecular characterization of a third blastoderm-specific locus, bsg25D, which maps to chromosomal locus 25D3. The bsg25D locus is defined as the DNA which encodes a 2.7 kb blastoderm-specific RNA and overlapping transcripts. We have determined the genomic DNA sequence of the bsg25D locus and the primary structure of the bsg25D RNAs, and have carried out computer database searches for homologies to the protein encoded by these RNAs.

MATERIALS AND METHODS

Unless otherwise noted, routine handling of nucleic acids followed standard protocols (5).

DNA sequencing

Both genomic DNA and cDNA were sequenced by the chain termination method (6) using buffer gradient gels (7). Most of the sequence was obtained from random subclones generated by sonication (8). Additional sequence was determined from subclones generated by digestion of large DNA fragments with four-cutter restriction enzymes, DNAase I digestion of large subclones (9), and digestion of large subclones with exonucleases III (10) and VII (11).

Isolation of cDNA clones

cDNA clones were isolated from two embryonic cDNA libraries (12, Goldschmidt-Clermont and Hogness, unpublished) by plaque hybridization (13).

Transcription mapping

Two microgram aliquots of poly(A)$^+$ RNA prepared from 1.5-3.5 hour embryos were electrophoresed and blotted as described (2). RNA was detected by a sandwich technique, in which small, single-stranded M13 probes (unlabeled) were first hybridized to the blot, followed by [^{32}P] nick-translated M13 RF DNA (14).

Primer extension and RNA sequencing were as described (15), using 6 or 12 micrograms of poly(A)$^+$ RNA from 1.5-3.5 hour embryos, respectively. Hybridization of 10^5-10^6 cpm of probe was for 18 hours at 52°C. Prior to sequencing, the hybridization mixture was divided into 4 equal aliquots.

S1 nuclease analysis was carried out essentially as described (16) using 10^5-10^6 cpm of probe and 6 micrograms of poly(A)$^+$ RNA from 1.5-3.5 hour embryos. Hybridizations were carried out for 18 hours at the temperatures indicated in the legend to Fig. 4, followed by digestion with S1 nuclease (500 units, BRL) for 1 hour at 37°C. Reaction products were electrophoresed on sequencing gels.

Computer analysis

The DNA sequence was compiled using the DB system (17,18). Codon usage analysis and translation were conducted using the ANALYSEQ package (19). The standard codon frequency table for this analysis was compiled from 20 *Drosophila* protein coding genes (20).

Searching the National Biomedical Research Foundation (NBRF) protein database was conducted using both the LSRCHP program (21) and the SEARCH program distributed by the Protein Information Resource (PIR; 22). Potentially homologous sequences were aligned by the ALIGN program, also distributed by the PIR. Probabilities that alignment scores would occur due to chance alone were calculated based upon the normal distribution, since random scores generated by the ALIGN program follow this distribution (23). The probability that an 8 amino acid identity would occur due to chance alone was calculated according to Kabsch and Sander (24), using an average frequency of occurrence for each of these 8 amino acids of 0.062 based upon data reviewed in (25).

Hydrophobicity correlation coefficients were calculated according to Sweet and Eisenberg (26) using each of the four hydrophobicity scales compared therein. One randomization of the *bsg25D* sequence produced by the ALIGN program was used as a control for unrelated structures (see Table 1).

RESULTS

Physical organization of the bsg25D locus

The genomic DNA clone containing the *bsg25D* locus, IB150, contains two Eco RI fragments of 9.1 and 7.0 kb which hybridize to four RNA species (2). The 9.1 kb Eco RI fragment hybridizes to a 4.4 kb transcript (Fig. 3, lane 9.1) which is expressed throughout most of embryogenesis (2). The 7.0 kb Eco RI fragment hybridizes to three RNAs of 2.7, 3.0 and 4.5 kb (Fig. 3, lane 7 and data not shown). The overlap of the latter RNAs is most clearly demonstrated by their hybridization to a single cDNA clone homologous to a portion of the 7.0 kb Eco RI fragment (Fig. 3, lane c3). The 4.5 and 3.0 kb RNAs are expressed primarily during the first 8 hrs of embryogenesis; the 2.7 kb transcript is blastoderm-specific (2).

The physical map of the *bsg25D* locus, which exists as a single copy in the haploid genome (20), is shown in Fig. 1. The 4.5 kb RNA overlaps with the 2.7 kb blastoderm-specific transcript, as indicated in Fig. 1A. This is

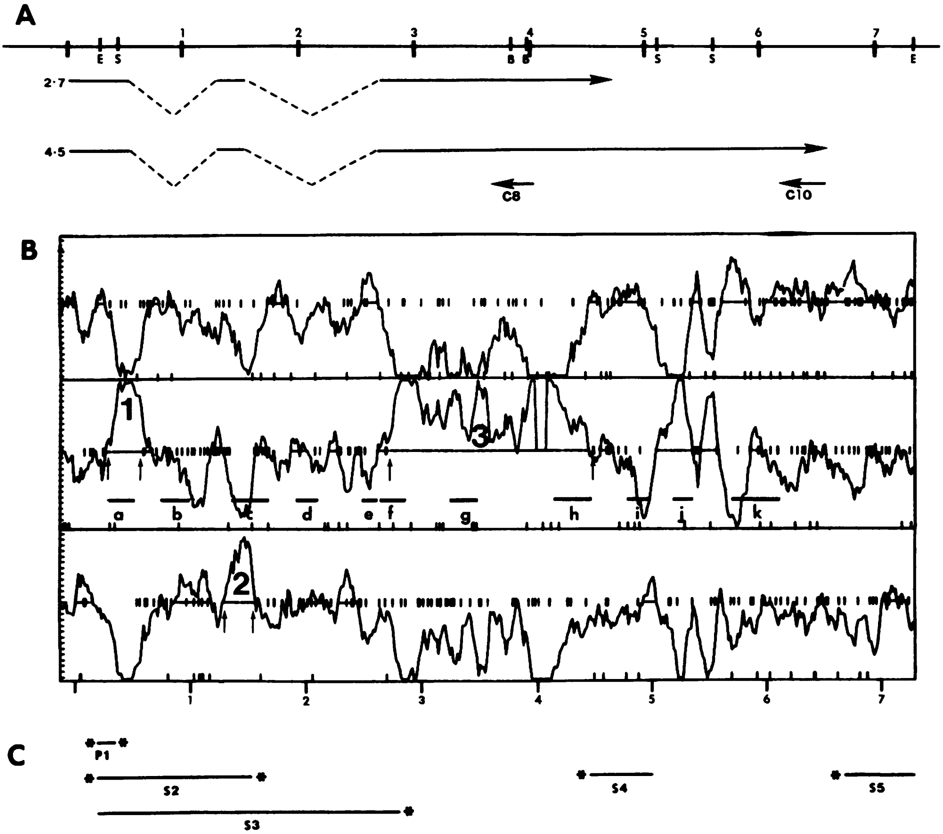

Figure 1. Organization of the bsg25D locus. A) Restriction map and transcription map. Top line represents the genomic DNA whose sequence is presented in Fig. 2. Numbers above ticks represent nucleotide positions in kilobases, letters below ticks represent restriction enzyme recognition sites (E: Eco RI; S: Sst I; B: Bgl II). The lower two lines labeled 2.7 and 4.5 represent the transcription map of the RNAs of these sizes; solid lines are exons, hatched lines introns. The two small arrows labeled c8 and c10 represent the map positions of partial cDNA sequences determined from clones cDNA-8 and cDNA-10. B) Codon usage analysis of the bsg25D nucleotide sequence. Probabilities for coding are plotted along the vertical axis (the mid-height position of each panel represents a probability of 50%) and nucleotide position is plotted along the horizontal axis in kilobases. Each of the three panels represents one of the three possible reading frames. Ticks at mid-height represent stop codons and ticks at the bottom of each panel represent AUG codons. Open reading frames included in the bsg25D RNAs are numbered 1, 2, and 3 and each is demarcated by upward arrows. Single-stranded probes used for the experiments shown in Fig. 3 are shown in the middle panel and represent nucleotides: 275-484 (a); 796-1008 (b); 1409-1597 (c); 1921-2093 (d); 2495-2639 (e); 2672-2894 (f); 3347-3523 (g); 4249-4529 (h); 4904-5075 (i); 5273-5433 (j); and 5787-6187 (k; the 3' end of this probe is approximate since it was determined from an agarose gel). C) Probes used

for transcription mapping experiments represent nucleotides: 276-426 (P1); 276-1589 (S2); 276-2836 (S3); 4460-4987 (S4); and 6637-7345 (S5). Asterisks represent the end which was labeled in these experiments; asterisks on both ends of a probe signify that the probe was uniformly labeled.

```
 -120                 -100                  -80                   -60                   -40
atcaatctaa cgatagtgta taacgatagg aacaatggtc cacgatatgg ccacctccgt gcaagtttgc ttaatgccct ccagagcgcg ccaccgtgct
 -20                  1                     20                    40                    60
cgctatactg cattaattgt tttttATCAA CTCGCTAGAA ATACGCTATC CCAAAAAACC GCAAACCCGC GATGTTTATG TTGCGTTCCG AAGTGCATAT
 80                   100                   120                   140                   160
CATAGATTAG TAGTAGTAGT AACCCCTCAA ACAGCCTGCT GTCCAAAAAA CACGCGTGAT TCCCCCGCCA CCCACGCACA TAGACCCCGA TATTTCACTT
 180                  200                   220                   240                   260
TTCTTGTTTT CGACCCCTGA CTGCGTTTGT GGATTTTCCC CCCAAGAAAA AAAAAGCGAA GTGAAAACGC AATTGAGCAG CCGATCGATT GGAACGGCAG
 280                  300                   320                   340                   360
GAATTCCCCG GGTTACGGAT AATGGAGGTA TCCGCCGATC CGTACGAGCA GAAGCTCTAC CAAATGTTCC GCAGCTGCGA GACGCAGTGT GGACTTCTGG
                        M  E  V   S  A  D   P  Y  E  Q   K  L  Y   Q  M  F   R  S  C  E   T  Q  C   G  L  L D
 380                  400                   420                   440                   460
ACGAGAAGTC CCTGCTGAAG CTCTGCTCAC TGCTGGAGCT CCGGGATCAG GGATCCGCAC TGATCGCCAG CCTGGGCGGC AGCCATCAGC TGGGCGTGTC
   E  K  S   L  L  K   L  C  S   L  L  E  L   R  D  Q   G  S  A   L  I  A  S   L  G  G   S  H  Q   L  G  V  S
 480                  500                   520                   540                   560
CTTTGGCCAG TTCAAGGAGG CGCTACTCAA CTTCCTGGGC TCCGAGTTCG ATGgtaatac gtcatcgggt ttcattggtg agatagcaca aagaatcgat
  F  G  Q   F  K  E   A  L  L  N   F  L  G   S  E  F   D  D
 580                  600                   620                   640                   660
cacgctatag attaacttat atagtataaa gataatattt gctataagct aacgcgacag gttcgcataa aacaacatac gttttatctg taattgcgct
 680                  700                   720                   740                   760
ttaattaccc atcaagcaac atcagataat tacggaatgt ttgccagcca cttattagag atagtaattc aattttgaca cggatttgga accgtgtggg
 780                  800                   820                   840                   860
tttccctatt aataaaacac tgatctaatg aacacatttc tagcagtcta tagatgaaca aagccattac ttaatactca aagaagtgct accatctacg
 880                  900                   920                   940                   960
tgctaatttg caaggattat gcacatttac ttcaaacctc cgcttatctg atttggaaac ttctgggcaa atttaggaca ccttagggta cgaatatcat
 980                  1000                  1020                  1040                  1060
aatcagcacg cggattagca cgcggcagct ggcgatcata aaatcataga tgcaattgac acttttttac gactcccaac tgttctcgac tacctgatcc
 1080                 1100                  1120                  1140                  1160
tgcatgatcc ttatcaggta gatggttaca atgtcctgta taaatacgcg acacattcac ctgggcagtt tagtctaaat caaaatggga acacgattgt
 1180                 1200                  1220                  1240                  1260
attaccgccg atccggcggt cagttaacag atccgataat tgagaagcta gccgctcgtt ttggtagcca cctaagatcc atacaactct tccagttctc
 1280                 1300                  1320                  1340                  1360
tgctaactta tatctattga atcttccagA GCGTTCACTG GTGATTACGG ATGAGCCGCT AAACAACACA TACATCGAGA GTCCGCCGGA GTCTTCCGAT
                                   R  S  L   V  I  T   D  E  P  L   N  N  T   Y  I  E   S  P  P  E   S  S  D
 1380                 1400                  1420                  1440                  1460
CGCGAGGTTT CACCCAAACT CGTCGTGGGC ACCAAGAAAT ACGGTCGCCG GTCTAGGCCA CAGCAGGGAA TCTACGAGTT ATCCGTCACG GACTCGGACA
 R  E  V   S  P  K  L   V  V  G   T  K  K   Y  G  R  R   S  R  P   Q  Q  G   I  Y  E  L   S  V  T   D  S  D N
 1480                 1500                  1520                  1540                  1560
ATACGGACGA GGACCAGTTG CAGCAGCAGC AAAATCAGCG AAGCCTCAAC GGATGCGATG AGCTGGGAGT TCAGgtgagt gtcgtttgtc aagtcacgta
   T  D  E   D  Q  L   Q  Q  Q   Q  N  Q  R   S  L  N   G  C  D   E  L  G  V   Q
 1580                 1600                  1620                  1640                  1660
cgaagtggcg atacaacttc tggtatgtat gcaaaattgc atagtaaaca gattttgttt aatcgttatt attgctgata cagtagagca tgcctaagta
 1680                 1700                  1720                  1740                  1760
gcactaccaa agcaaacaaa ttatcttaaa tatacatcat gatcatcata agcatcttat ttttccaaac cacacaggtg caacgttcct cgtcccagag
 1780                 1800                  1820                  1840                  1860
cgatcttcct ggcagccggc gtctgcggtc cgtccacacc agcgggagca aactgaagcg ttgtgcttca ctgccagccc gccggaagat gaacagcaac
 1880                 1900                  1920                  1940                  1960
accacggagc cactacatca ccgacggcag cggccaagtt gaaacagctt tccatccaga gccaggcgca gcacagcagc agcgtggaat cactgggtaa
 1980                 2000                  2020                  2040                  2060
gtttcctctg gccagaccag ctttggctag ccgatccccc ttgtccctgc caccctctgt tgttgttagc ccaaaatgcc aaaattacgt ttgaagcaat
 2080                 2100                  2120                  2140                  2160
gttaaaagca aaacacttgt ttgtcggtac acaccgtagc catcgcctgg ccaccaatcc cgcaccgtcg tccgagcact ggagatgcta ccacggcggc
 2180                 2200                  2220                  2240                  2260
cgttggtcat gctgcaaagg tttgtgcgct ctgaagcaat tgtcaacacc ctcacaccca ccgaatcccc aacccagtca ttcggtatct aatcgcaccc
 2280                 2300                  2320                  2340                  2360
tatgtagccg cacatttgat tcgttttttt tactcgtata ataacatatc ctacattttc aacccttagt aatgctgtaa tgcattgaca atcaatttaa
 2380                 2400                  2420                  2440                  2460
ttaaggattt catataaatc aatttcagtt agaaaggata tttacttata atttgttcta ttttcttgat ttattagttt ctacctcttt aaataacacg
```

```
2480                  2500                  2520                  2540                  2560
gcaaaaattt ctcatttcta aaagccattt gatatagaga aataacaaac tttcggcgct tttgcttaca ccatcgacac acacacacac ccttccccac
2580                  2600                  2620                  2640                  2660
tcccaatccc aatccaatcc cacacccacc tggtatcttg ggctatatgt ataaaaatgt gtatatacaa cagcgaagcc aatctcattc gtcccacgct
2680                  2700                  2720                  2740                  2760
aattgttaat tgccatgatt tacagacacc gtgacgccgc agCAATTGGA GACGATCTCA GTGCATAGCA TTATGGAAGC CTGGGAGCTG GCCAGCATTC
                                              Q  L  E  T  I  S  V  H  S  I  M  E  A  W  E  L  A  S  I  P
2780                  2800                  2820                  2840                  2860
CCAACACTCG CAACCTACTT CACGTCCTGG GATTCGATGA GGAGGAGGAG GTGAACCTGC AGCAGCTAAC TAAGGCATTG GAGGAGGAGC TGCGGGGCAT
  N  T  R  N  L  L  H  V  L  G  F  D  E  E  E  E  V  N  L  Q  Q  L  T  K  A  L  E  E  E  L  R  G  I
2880                  2900                  2920                  2940                  2960
CGATGGGGAT CACGAGCAAT CGAATATGTT GCGCGCTCTG GCTGCTCTGC AGGCCACCGA GTTGGGCAAC TACAGACTTG CCTATAGGCA GCAGCATGAG
 D  G  D  H  E  Q  S  N  M  L  R  A  L  A  A  L  Q  A  T  E  L  G  N  Y  R  L  A  Y  R  Q  Q  H  E
2980                  3000                  3020                  3040                  3060
GAGAACCTCA AGCTGAGGGC CGATAATAAG GCGGCCAACC AAAGGGTGGC TTTGCTTGCC GTGGAAGTGG ATGAGCGGCA TGCGTCGCTG GAGGATAACT
E  N  L  K  L  R  A  D  N  K  A  A  N  Q  R  V  A  L  L  A  V  E  V  D  E  R  H  A  S  L  E  D  N  S
3080                  3100                  3120                  3140                  3160
CCAAGAAGCA GGTGCAGCAG CTGGAGCAAA GACACGCCAG CATGGTGCGT GAAATAACGC TGCGGATGAC TAATGACCGC GATCACTGGA CCAGCATGAC
  K  K  Q  V  Q  Q  L  E  Q  R  H  A  S  M  V  R  E  I  T  L  R  M  T  N  D  R  D  H  W  T  S  M  T
3180                  3200                  3220                  3240                  3260
GGGAAAGCTG GAGGCACAGC TTAAATCGCT TGAGCAGGAG GAGATCCGTC TGAGAACGGA ACTTGAACTG GTGCGCACTG AGAACACGGA GCTTGAGTCG
 G  K  L  E  A  Q  L  K  S  L  E  Q  E  E  I  R  L  R  T  E  L  E  L  V  R  T  E  N  T  E  L  E  S
3280                  3300                  3320                  3340                  3360
GAGCAGCAAA AGGCTCACAT CCAAATCACA GAGCTTCTCG AACAGAACAT TAAGCTCAAC CAGGAACTGG CCCAAAGGTC GAGCAGCATT GGTGGCACCC
E  Q  Q  K  A  H  I  Q  I  T  E  L  L  E  Q  N  I  K  L  N  Q  E  L  A  Q  R  S  S  S  I  G  G  T  P
3380                  3400                  3420                  3440                  3460
CGGAGCACAG TCCATTGCGA CCGAGAAGGC ATAGCGAGGA CAAGGAGGAG GAGATGCTCC AGCTAATGGA GAAGCTGGCT GCTCTTCAAA TGGAGAACGC
  E  H  S  P  L  R  P  R  R  H  S  E  D  K  E  E  E  M  L  Q  L  M  E  K  L  A  A  L  Q  M  E  N  A
3480                  3500                  3520                  3540                  3560
CCAGCTGCGT GACAAGACTG ACGAACTGAC CATCGAAATC GAGAGCTTAA ATGTGGAACT AATTCGCTCG AAAACCAAGG CTAAAAAGCA AGAAAAACAG
 Q  L  R  D  K  T  D  E  L  T  I  E  I  E  S  L  N  V  E  L  I  R  S  K  T  K  A  K  K  Q  E  K  Q
3580                  3600                  3620                  3640                  3660
GAGAAACAAG AGGACCAGGA GTCGGCGGCC ACGGCTACCA AAAGGCGTGG GGATTCGCCG AGCAAAACAC ATCTAACAGA GGAGAGCCCT CGCTTGGGGA
E  K  Q  E  D  Q  E  S  A  A  T  A  T  K  R  R  G  D  S  P  S  K  T  H  L  T  E  E  S  P  R  L  G  K
3680                  3700                  3720                  3740                  3760
AACAGCGCAA GTGCACCGAA GGAGAGCAGA GCGATGCCAG CAACAGCGGA GATTGGTTGG CTCTAAACTC CGAGCTGCAA AGAAGTCAAA GCCAGGATGA
  Q  R  K  C  T  E  G  E  Q  S  D  A  S  N  S  G  D  W  L  A  L  N  S  E  L  Q  R  S  Q  S  Q  D  E
3780                  3800                  3820                  3840                  3860
GGAGCTAACA AGCCTTAGAC AGCGGGTTGC TGAGCTAGAG GAGGAACTCA AGGCTGCAAA GGAAGGCAGA TCTCTCACCC CGGAAAGCCG TTCGAAGGAA
 E  L  T  S  L  R  Q  R  V  A  E  L  E  E  E  L  K  A  A  K  E  G  R  S  L  T  P  E  S  R  S  K  E
3880                  3900                  3920                  3940                  3960
CTGGAGACCA GTCTAGAGCA AATGCAGCGT GCCTATGAGG ATTGCGAGGA CTACTGGCAA ACGAAACTTA GCGAGGAGCG GCAGCTGTTT GAGAAGGAGC
L  E  T  S  L  E  Q  M  Q  R  A  Y  E  D  C  E  D  Y  W  Q  T  K  L  S  E  E  R  Q  L  F  E  K  E  R
3980                  4000                  4020                  4040                  4060
GACAGATCTA CGAAGATGAG CAGCACGAGA GCGACAAGAA GTTCACCGAG CTGATGGAAA AGGTGCGCGA GTACGAGGAG CAGTTCAGCA AGGATGGCCG
  Q  I  Y  E  D  E  Q  H  E  S  D  K  K  F  T  E  L  M  E  K  V  R  E  Y  E  E  Q  F  S  K  D  G  R
4080     ←8           4100                  4120                  4140                  4160
CCTCTCGCCC ATTGATGAGC GCGATATGCT GGAACAGCAG TACTCGGAAT TGGAGGCAGA GGCAGCCCAG CTGCGCTCGA GTTCCATTCA AATGCTCGAG
 L  S  P  I  D  E  R  D  M  L  E  Q  Q  Y  S  E  L  E  A  E  A  A  Q  L  R  S  S  S  I  Q  M  L  E
4180                  4200                  4220                  4240                  4260
GAGAAGGCTC AGGAAATCAG CTCACTGCAA TCGGAGATCG AGGATTTGCG ACAGAGATTG GGTGAGAGCG TTGAGATCCT TACAGGCGCC TGTGAACTCA
E  K  A  Q  E  I  S  S  L  Q  S  E  I  E  D  L  R  Q  R  L  G  E  S  V  E  I  L  T  G  A  C  E  L  T
4280                  4300                  4320                  4340                  4360
CCTCGGAGTC GGTAGCCCAA CTGAGTGCCG AGGCGGGAAA AAGTCCAGCC AGCTCACCCA TCAGCTACCT CTGGCTGCAG AGCACCATCC AAGAGCCAGC
  S  E  S  V  A  Q  L  S  A  E  A  G  K  S  P  A  S  S  P  I  S  Y  L  W  L  Q  S  T  I  Q  E  P  A
4380                  4400                  4420                  4440                  4460
GAAATCGCTT GCCGATTCCA AGGATGAAGC CACCGCCAGT GCCATCGAAT TGCTCGGAGG CTCACCATCG CACAAGACAG CCAGCCGGTG AGTATGAGAA
 K  S  L  A  D  S  K  D  E  A  T  A  S  A  I  E  L  L  G  G  S  P  S  H  K  T  A  S  R  *
4480                  4500                  4520                  4540                  4560
GCCTCTCGGT GTGTCCTTGG TGTGAGCATC CCTGTGTCTT CCTCATAATT TGCACTGTAT GTCCTGTATA TATGTTTCAG TTTGTCCCTC ACATCTAACC
4580                  4600                  4620                  4640                  4660
ATGTCTAATA TAAGCTAATT TAATCCTTTT AATTGTATGT TTGTGCTTGT TTAATAAATA TAATTTATAT TCATATAGAA ATTCATCACA TTATCGAAAT
              ↑                  ↑                    ↑                                              ↑
```

```
4680                  4700                  4720                  4740                  4760
TCATTGATTT ATGATTTCAA TAAATATACA TTTAATATTT TACAAAAAAA TTACTCTTTT TCGGTTATGA AATTGGCTGC TGGAAATGGT TTTGTTTGCT
4780                  4800       ↑          4820↖                 4840                  4860
TATTTTTCAC ATTTGTATCA TTACACGTTT TGCATCTTTA TGTTACATCT TCAATCGTTT TTATTTTGTA AATCATGCCA TTTAATGGTC CCTTAAACAG
4880                  4900                  4920                  4940                  4960
CAATAACCTC ACCACTTCGG AAACATCCAT CTTTAGCACT ACACCCTTCG AAAGCTCTCA GTCGGGTCCT TCGCCCACGA ACAGTGGCAA CAGCAACGCC
4980                  5000                  5020                  5040                  5060
TACGGCCAAT CCCGGCCCAG CTCCGATCAG CAAGCCCAAG CGGTCCCAGA GTCCCCAACA GGCGGCTGCA TCGGAGGGAG AGATAGCCGA TTGCGAGACG
5080                  5100                  5120                  5140                  5160
TCGTCGACGG CGTCCGGCAA AAGCTTCGAA TCCAACAGTA AAACGTCTTG CCTTAGCCAC GAGAAGTGCA GCAGTCCGTC GGCACTGAAG GAGGAACTGA
5180                  5200                  5220                  5240                  5260
AGCGCCTTAA GTTCTTCGAG CTCTCCCTCA AGGAGCAAAT CAAGGATCTG AGTCTGCAGC GGGACGGTCT GGTCATGGAA CTGCAGCAGT TGCAGGAGGC
5280                  5300                  5320                  5340                  5360
GCGACCCGTG CTCGAGAAGG CCTATGCGGT GAGCCATAGA CTTTGTTGAT CAGGGAACAT ATTCTAATCG TATCTGTGGA CTCTTCTTTA GCGAACAACG
5380                  5400                  5420                  5440                  5460
CATCCAACGC TTCAGCAGCG ACTGAACCAA TTGGAGCTGC GAAATCGCCA TCTGCAGAAT GTCATCAAGC AGCAGCAGCA TTACACGGAG TCCCTGATGC
5480                  5500                  5520                  5540                  5560
AGCGTAAGTT GAAAAACTAC CTACTGACCA AGAATGATAA TGTAATATTT ATTCATTAGA ATCCTGGCGG CAGCATCAAG TGGAGCTCAA CGATTTGCAT
5580                  5600                  5620                  5640                  5660
AGCCGAATCG AGACCAGGGT GTTTTACTGG CCGATCAGAC ACAGCGATTG CAGAGTGCCG ACATCCTGGT GAAGGATCTA TATGTGGAGA ACTCCCATCT
5680                  5700                  5720                  5740                  5760
GACGGCCACG GTGCAGCGGT TGGAGCAGCA ACGAGCTAGG GTGAACCTCA TTCACCAGCA GCAGCAACAG CAGCGCCTTG TGGGCGGTGG ACTGCCTGGC
5780                  5800                  5820                  5840                  5860
ATGCCTTAGT TTGCCCCACC GGCAAACGTA TATAGTTTAT AGATAATTAT GAAAAAGACA AACCTGAGGA GGGAGTGGTG CTCAGCATCG GCAGACATCG
5880                  5900                  5920                  5940                  5960
AACATGCACC TAACCATAGA TCCTTATGAA TGTTTAGACA TATACAATTC TCGGTAGATT AAGTTTGCAT ACCCGTCGTA TTCGTATTCG TACGTTGCGT
5980                  6000                  6020                  6040                  6060
TTTTTTTGTG AATGAATGTG AATCCATGTT GTTCGACACG AGAGCACAGC AGCAATAACT AAAGTGACTT TAAACTAAAC TTAAACTCAC CCACGCGCAA
6080                  6100                  6120                  6140                  6160
ATGAGGAACA ATCCACACTA GTGTACCAAT TTGTAACACA TCTAGTAATC GAATCGACTA AACTATTTAC ACGAGCTACA GGACATATAC GATGAAGTAC
6180                  6200                  6220                  6240                  6260
CCACGTAGTA TATGTTCGTG CAATGTTGAC CTTACTAATT GACTACTGAA ACAGTTATCG TATATTAATT ATATTAGAAG AAACAGTATT TTAAATTTGT
6280                  6300                  6320                  6340                  6360
TATGCGTCTG AGTAGGCGAG CACGTTTATC AATGTTTATC ACGTGCCCAA TCAAATGCAT CGGAATTGTT GTTAATTTTA TTGATAGAGA AAATGGAAAT
6380                  6400                  6420                  6440                  6460
GAGCGTAAAA AATGATCTAT GATATTGATA TTGATGTAAT ATTTAACGAC AAAAGACCTG TAAAGCTGTA ACCATACACA CGAATCTATG TATTTAAATT
6480                  6500                  6520                  6540                  6560
GCGATCTAAG TTAGCCAATA CTCTTCAATA TTGCTTTTGC GAACGCGACT TTTTGTTATA TCTTCATTCG TCCCAATAAC TCACTCGATT TATATGTAAA
6580                  6600                  6620                  6640                  6660
GAAAAAAAAA CTCAACCTCA ATCACACGAT ATCGTGTAAT CAGTGCTTAA ATCAATACTT TCGATCAAAA TAGAAGTTTA CTTTTTAAAA GTATAAAAAA
6680          10      6700                  6720                  6740                  6760
            ←
TAATACAACA AAAAAACCAA ATatacaatt atttataaaa ccaaattgtg ataactcgtc tttattctaa atagttatta aaatgttgcg ggaatataaa
6780                 ↗6800                  6820                  6840                  6860
cttattgttc ataaatacaa cttgcttatc agtttttttgg aaatgttaag attttgtttc ttattaaatt aattgttatt aaaattaaaa gatttatgaa
6880                  6900                  6920                  6940                  6960
gtttaaatta tatattgata cgataaacaa tttattttat tgcttcaaaa tatacattac tttttttttag gaatttaaat atccgtttta agtcttttaa
6980                  7000                  7020                  7040                  7060
tttttaatta ggttttaaat ataatatcga ttaaatagtt gactccattg gaatatcgat accgcgtcga tgtttcttcc agctctatcg ggcacgcgct
7080                  7100                  7120                  7140                  7160
gttaaagttt atttgtactg ttaacgcgaa ttggattaaa atgttttgtt ttttgttagt tttcggtgta aatgtggttt tagtgcactt aacaactggt
7180                  7200                  7220
gagtgtgcgt tatagtaaat gaacttaaat gcaattaccg aattc
```

Figure 2. Nucleotide sequence of the bsg25D locus. The sequence is shown below numbers which indicate nucleotide position, with transcription initiation at nt 1. The TATA homology is underlined; the transcription initiation site is indicated by an arrow; exons are indicated by upper case letters; introns are indicated by lower case letters; 3' ends of partial cDNA sequences (from nt 3724-4089 for cDNA-8 and from nt 6259-6693 for cDNA-10) are indicated by leftward arrows below numbers corresponding to the cDNA clone; poly(A) addition signals discussed in the text are underlined; and 3' ends of transcripts are indicated by upward arrows.

also the case for the 3.0 kb RNA (see below), but it is too rare to be mapped by available methods.

The complete nucleotide sequence of the bsg25D locus is presented in Fig. 2. This sequence includes 125 nt upstream of the transcription initiation site, three exons and two introns, and 523 nt downstream of the 3' end of the 4.5 kb RNA (see next section). Approximately 70% of this DNA was sequenced on both strands. The accuracy of the remaining DNA sequence was insured by sequencing multiple clones representing these regions, and ambiguities were resolved by substituting dITP for dGTP in the sequencing reactions (20).

Transcription mapping the bsg25D RNAs

The results of transcription mapping show that the 2.7 and 4.5 kb bsg25D RNAs initiate at the same site, that they have two intervening sequences which are spliced in the same positions, and that the greater length of the 4.5 kb RNA results from read-through transcription past the 3' terminus of the 2.7 kb RNA (Fig. 1A). Five independent lines of evidence support this conclusion:

1) Nine cDNA clones analyzed (out of 31 isolated, see Materials and Methods) fell into two classes: one which hybridized to both the 2.7 kb and 4.5 kb RNAs (8/9), and one which hybridized almost exclusively to the 4.5 kb RNA (1/9) (Fig. 3, lanes c3 and c10). Sequence analysis allowed mapping of the 3' end of two cDNA inserts, one from each class, to the positions shown in Fig. 1A.

2) Codon usage analysis (19) of the bsg25D DNA sequence suggests the presence of three open reading frames with high probabilities for protein coding (Fig. 1B); these open reading frames correspond closely with the proposed 2.7 kb RNA exons (Fig. 1A).

3) Hybridization of small single-stranded probes to RNA gel blots indicates that the two RNAs overlap, that they share three exons, and that the 4.5 kb RNA is derived from a region beyond the 3' end of the 2.7 kb RNA (Fig. 3a-k and legend).

4) RNA endpoints, determined by primer extension and S1 nuclease analysis (Fig. 4), are consistent with the proposed transcription map. RNA sequencing by primer extension is collinear with the DNA sequence to nt 39 (as far as the sequence could be read), consistent with initiation at nt 1. The precise positions of the 5' and 3' ends of the three exons, determined by S1 mapping, are summarized in the legend to Fig. 4. The terminus of the 4.5

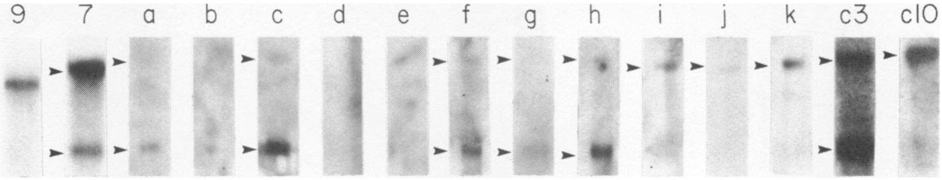

Figure 3. Mapping bsg25D RNAs by probing RNA gel blots. Arrows indicate hybridization to the 4.5 and 2.7 kb bsg25D RNAs. Blots of 1.5-3.5 hr poly(A)$^+$ RNA were hybridized as follows. Lanes 9 and 7 were probed with ^{32}P nick-translated 9 and 7 kb Eco RI fragments from clone IB150. Lanes c3 and c10 were probed with cDNA-3 or cDNA-10, which were nick-translated with ^{32}P. Lanes a-k were hybridized with the small single-stranded DNA fragments a-k shown in Fig 1B. The small size of these probes and the rarity of the bsg25D RNAs required the design of a novel hybridization protocol (14) in order to detect the low signal shown in these experiments. Probes from regions encoding exons hybridize to the 2.7 kb RNA (lanes a,c,f,g,h), while those from regions encoding introns do not (lanes b,d,e). In addition, all probes which hybridize to the 2.7 kb RNA also hybridize to the 4.5 kb RNA, although this is difficult to detect in the photographic reproductions of some lanes. Probes from the region encoding the 3' end of the third exon hybridize only to the 4.5 kb RNA (lanes i,j,k).

kb RNA deduced by S1 analysis is consistent with the location of the 3' end of cDNA-10 at nucleotide 6692.

5) Transcriptional signals in the DNA sequence (Fig. 2) are consistent with the RNA endpoints determined in the preceding experiments. Upstream from the transcription initiation site is a TATA sequence, as is usually found for genes transcribed by RNA polymerase II (28). The transcription initiation site is homologous to other Drosophila initiation sites (not shown, 15). Sequences at the splice junctions between exons 1, 2, and 3 are all reasonably homologous to consensus splice junction sequences (27), and in each case the GT-AG splicing rule (29) is strictly followed and the open reading frames are joined in frame. There are two consensus poly(A) addition signals at positions 4628 and 4694. The last three of the endpoints determined by S1 nuclease analysis for the 2.7 kb RNA are consistent with recognition of these signals, but the first three endpoints are not preceded by similar signals (see Discussion). Although there are no consensus poly(A) addition signals located near the designated 3' end of the 4.5 kb RNA, two variants of this sequence located at nucleotides 6667 and 6677; both of these have been shown to be functional poly(A) additional signals in other systems (see Discussion).

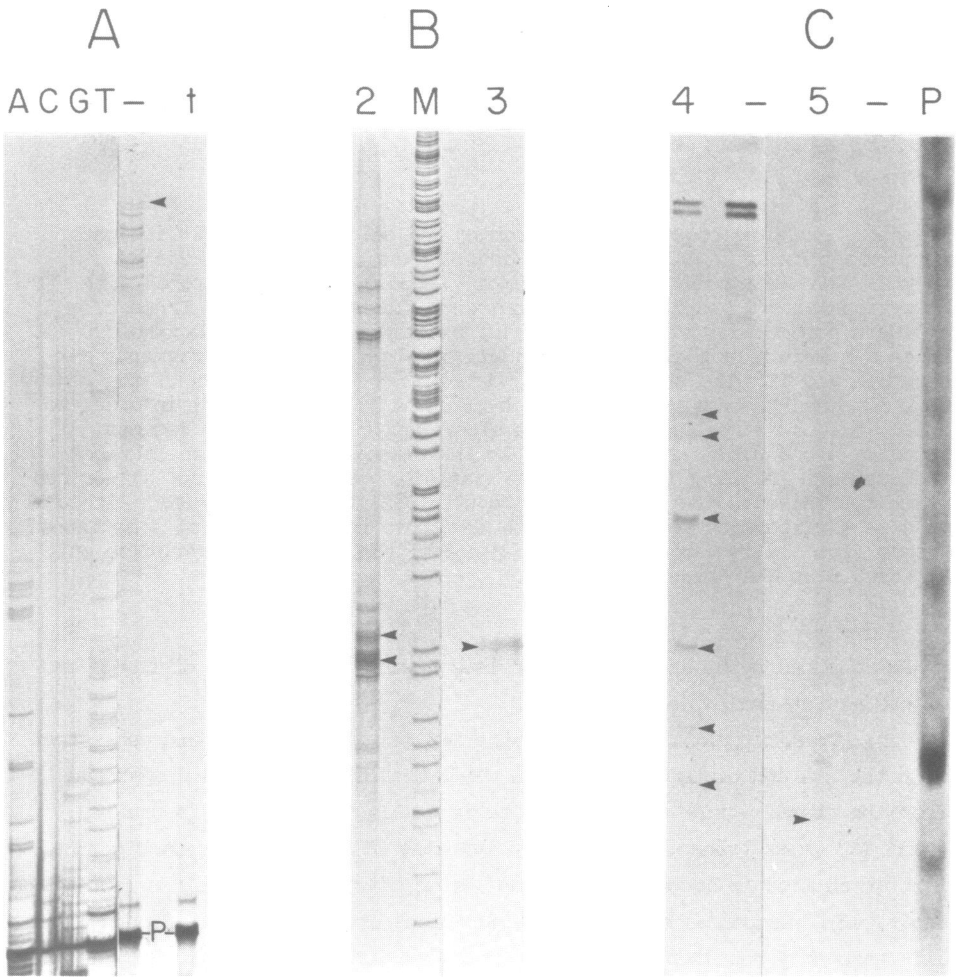

Figure 4. Mapping bsg25D RNAs by S1 and primer extension. A) Determination of the 5' end. Uniformly labeled 150 nt primer P1 (Fig. 1C) was extended in the presence (lanes A, C, G, and T) or absence (lane –) of dideoxynucleotides. The arrowhead (lane –) indicates the longest product (275 nt) which places the transcription initiation site at nucleotide 1 (Fig. 2). The sequence of the RNA, determined using the same probe in the presence of dideoxynucleotides (lanes A, C, G, T) was collinear with the DNA sequence to nt 39 (as far as it could be read). Lane t contains products of a control reaction, where tRNA was substituted for blastoderm RNA. B) Mapping splice junctions. The lengths of exons 1 and 2 were determined using uniformly labeled probe S2 (Fig. 1C) which was hybridized to RNA at 52°C, followed by S1 digestion. Two distributions of protected fragments, centered around intense bands of 252 and 245 nt (arrows, lane 2), were found [other fragments in lane 2 were also present in control reactions which lacked blastoderm RNA (not shown)]. As preliminary S1 analysis (not shown) placed the 5' end of

exon 2 at position 1305, and a consensus donor splice junction sequence (27) sequence is located at position 1549, we assign the 245 nt protected fragment to exon 2. Assignment of the 252 nt protected fragment to exon 1 places the 3' end of this exon at nucleotide 528 (Fig. 2). Lane M is a sequencing lane representative of the standards used in all mapping experiments. The 5' end of exon 3 was determined by S1 analysis using probe S3, which was hybridized to RNA at 50°C. The arrowhead (lane 3) points to 114 and 115 nt protected fragments. On the basis of these fragment sizes and the consensus splice sequence, we place the 5' end of exon 3 at nucleotide 2717 (Fig. 2).
C) Determination of 3' ends of the 2.7 and 4.5 kb RNAs. Probe S4, hybridized to RNA at 43°C, was used to determine the 3' end of the 2.7 kb RNA. Arrowheads (lane 4) point to protected fragments of 128, 143, 163, 204, 244, and 258 nt; these fragment sizes indicate that the 3' ends of the 2.7 kb RNA are at nucleotides 4589, 4605, 4625, 4666, 4706, and 4720 (Fig. 2). Probe S5, hybridized to RNA at 30°C, was used to determine the 3' end of the 4.5 kb RNA. Only a very faint protected band was observed which was not present in control lanes (arrow in lane 5); this band was difficult to reproduce photographically. The position of the 3' end of the 4.5 kb RNA corresponding to this protected fragment is nucleotide 6697. Both - lanes in this panel are control S1 reactions for the respective probe in which tRNA was substituted for blastoderm RNA. Lane P in this panel is probe 5 which was not treated with S1 to show that the protected fragment is not present.

The total length of the exons mapped in the above experiments are 2.7 and 4.7 kb, consistent with the 2.7 and 4.5 kb sizes determined from RNA gel blots (2).

Protein database searches

The 741 amino acid sequence predicted from the nucleotide sequence is translated in Fig. 2. Codon usage analysis in Fig. 1B shows a probability for coding of at least 50% for the entire length of this amino acid sequence (see areas between arrows in Fig. 1B). The predicted bsg25D amino acid sequence was used to search the NBRF protein sequence database (see Materials and Methods). We discuss below the two highest scoring similarities. To avoid the functional and evolutionary implications associated with the term "homology", we use the term "similarity" to indicate a relationship identified by statistical analysis.

The SEARCH program identified a 96 amino acid domain with 22% identity between the bsg25D amino acid sequence and the product of the fos oncogene; one gap of two amino acids was inserted by the ALIGN program into the bsg25D sequence to optimize the alignment (Fig. 5). Also shown in this figure are regions of other gene products homologous to fos (see Discussion). The alignment score for the similarity to v-fos is 7.72 standard deviations above the average score for 100 randomizations of the respective sequences. The probability that this score could arise due to chance alone was calculated to be ~10^{-13} (see Materials and Methods). When this score is corrected for the

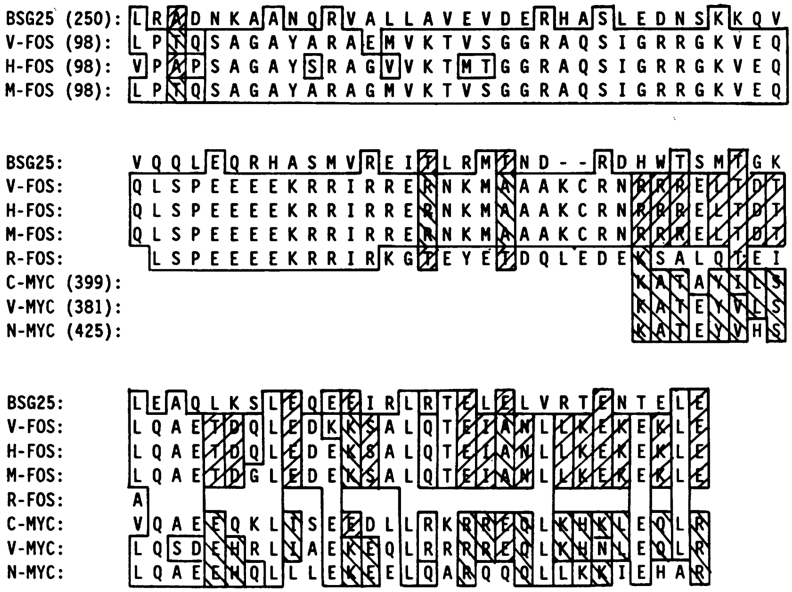
BSG25 (250): L R A D N K A A N Q R V A L L A V E V D E R H A S L E D N S K K Q V
V-FOS (98): L P T Q S A G A Y A R A E M V K T V S G G R A Q S I G R R G K V E Q
H-FOS (98): V P A P S A G A Y S R A G V V K T M T G G R A Q S I G R R G K V E Q
M-FOS (98): L P T Q S A G A Y A R A G M V K T V S G G R A Q S I G R R G K V E Q
BSG25: V Q Q L E Q R H A S M V R E I T L R M T N D - - R D H W T S M T G K
V-FOS: Q L S P E E E E K R R I R R E R N K M A A A K C R N R R R E L T D T
H-FOS: Q L S P E E E E K R R I R R E R N K M A A A K C R N R R R E L T D T
M-FOS: Q L S P E E E E K R R I R R E R N K M A A A K C R N R R R E L T D T
R-FOS: L S P E E E E K R R I R K G T E Y E T D Q L E D E K S A L Q T E I
C-MYC (399): K A T A Y I L S
V-MYC (381): K A T E Y V L S
N-MYC (425): K A T E Y V H S
BSG25: L E A Q L K S L E Q E E I R L R T E L E L V R T E N T E L E
V-FOS: L Q A E T D Q L E D K K S A L Q T E I A N L L K E K E K L E
H-FOS: L Q A E T D Q L E D E K S A L Q T E I A N L L K E K E K L E
M-FOS: L Q A E T D G L E D E K S A L Q T E I A N L L K E K E K L E
R-FOS: A
C-MYC: V Q A E E Q K L I S E E D L L R K R R E Q L K H K L E Q L R
V-MYC: L Q S D E H R L I A E K E Q L R R R R E Q L K H N L E Q L R
N-MYC: L Q A E E H Q L L L E K E E L Q A R Q Q Q L L K K I E H A R

Figure 5. Alignment of similar domains in the bsg25D and fos proteins. Residues 250-344 of the bsg25D amino acid sequence (BSG25) were aligned with residues 98-194 of the FBJ murine osteosarcoma virus (V-FOS) protein and cellular homologs from human (H-FOS) and mouse (M-FOS) by the align program. A penalty of 12 points was deducted from the raw score for the two-residue gap inserted in the bsg25D sequence. Regions of the r-fos amino acid sequence and myc homologs were aligned with the fos amino acid sequence as reported (36, 37). Two residues of the r-fos sequence between residues 8 and 9 in this figure were deleted (36) to optimize alignment with other fos sequences. Boxes enclose residues which are identical in two or more sequences. In cases where one amino acid is present in several proteins and another amino acid is present in several other proteins at the same position, boxes are hatched in opposite directions.

number of comparisons made in the database search, the probability that the alignment is due to chance alone is ~10^{-9}.

This similarity suggests that the two domains of the bsg25D and fos proteins may be folded in the same manner; we evaluated this by calculating the hydrophobicity correlation coefficient (26). As proteins with similar three-dimensional structures are characterized by coefficients of 0.3-0.7 (26), our calculation of 0.56 for the hydrophobicity correlation coefficient between the bsg25D and fos domains (Table I) suggests that these domains share a common three-dimensional structure.

The second similarity, identified by the LSRCHP program, is between a 21 amino acid segment of the bsg25D protein and repeated segments of tropomyosin

Table 1. Hydrophobicity correlation coefficients for the bsg25D-fos similarity domain[a].

	BSG25D	V-FOS	C-FOS	Ran-BSG[b]
BSG25D		0.555	0.560	0.229
V-FOS	0.555		0.996	0.227
C-FOS	0.560	0.996		0.229
Ran-BSG	0.229	0.227	0.229	

[a] Calculations were carried out as described in Materials and Methods using the consensus hydrophobicity scale of Sweet and Eisenberg (26). Similar values are obtained (data not shown) when the hydrophobicity scales of Dayhoff et al. (38), Wolfenden et al. (39,40), and Janin (41) are used.
[b] A randomization of the bsg25D-fos similarity domain sequence (see Materials and Methods).

(Fig. 6). The repeated tropomyosin segments contain characteristic clusters of negatively and positively charged residues; each alpha-helical segment is thought to bind a monomer of F-actin (30). Both the primary sequence of several of these repeated segments (shown by alignment scores greater than 3.0 in the right column) and the distribution of charged residues are shared by the similar segment in the bsg25D protein. In addition, one tropomyosin segment shares eight consecutive identical amino acids with the bsg25D segment. The probability that these eight identical residues would occur in two proteins due to chance alone was calculated to be ~10^{-9} (24). The occurrence of eight consecutive identical residues is further support for a structural relationship between the segments.

DISCUSSION

We have determined the complete nucleotide sequence of the bsg25D locus, as well as a transcription map supported by five independent lines of evidence (Fig. 1). The primary structure of the 2.7 kb RNA was used to deduce the amino acid sequence of the bsg25D protein. Database homology searches reveal two domains of the bsg25D protein which show structural similarity to domains of products of the fos oncogene and of tropomyosin.

The transcription map of the bsg25D locus raises several interesting issues. First, there are multiple 3' termini for the 2.7 kb RNA. Three of the six protected fragments are consistent with recognition of consensus poly(A) addition signals, but the three shorter protected fragments are not. These latter three fragments could result from "breathing" of the DNA-RNA hybrids during the S1 digestion; alternatively, RNAs of several different sizes could arise from recognition of variant poly(A) addition signals in an

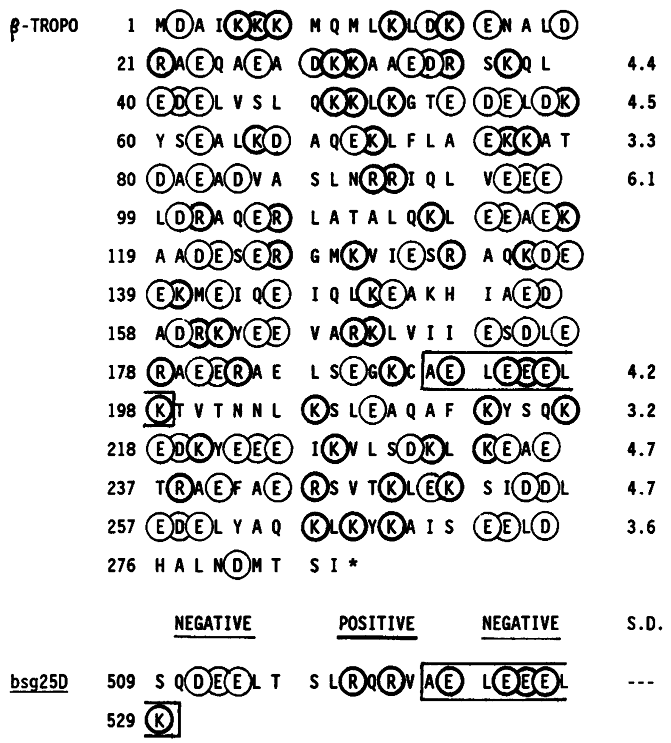
β-TROPO
1 M D A I K K K M Q M L K L D K E N A L D
21 R A E Q A E A D K K A A E D R S K Q L 4.4
40 E D E L V S L Q K K L K G T E D E L D K 4.5
60 Y S E A L K D A Q E K L F L A E K K A T 3.3
80 D A E A D V A S L N R R I Q L V E E E 6.1
99 L D R A Q E R L A T A L Q K L E E A E K
119 A A D E S E R G M K V I E S R A Q K D E
139 E K M E I Q E I Q L K E A K H I A E D
158 A D R K Y E E V A R K L V I I E S D L E
178 R A E E R A E L S E G K C A E L E E E L 4.2
198 K T V T N N L K S L E A Q A F K Y S Q K 3.2
218 E D K Y E E E I K V L S D K L K E A E 4.7
237 T R A E F A E R S V T K L E K S I D D L 4.7
257 E D E L Y A Q K L K Y K A I S E E L D 3.6
276 H A L N D M T S I *
NEGATIVE
POSITIVE
NEGATIVE
S.D.
bsg25D
509 S Q D E E L T S L R Q R V A E L E E E L ---
529 K

Figure 6. Similarity between a segment of the bsg25D protein and repeated segments of tropomyosin. The complete amino acid sequence of rabbit beta-tropomyosin is shown above the sequence of the bsg25D segment. Each line of tropomyosin sequence represents one proposed actin-binding domain (redrawn from 30). Within each domain, subdomains with concentrations of negatively and positively charged residues (shown by light and heavy circles) have been separated. Numbers on the left indicate positions in the respective amino acid sequences, while numbers on the right are alignment scores generated when this bsg25D segment is compared to the respective tropomyosin segment. The 8 amino acid identity in the two proteins is boxed.

AT-rich region. Second, variant poly(A) addition signals may also determine the endpoint of the 4.5 kb RNA, which is not located downstream of consensus poly(A) addition signals, but is a reasonable distance from two variants of the consensus sequence, both of which have been demonstrated to be functional in other systems (31, 32). Third, that both RNAs appear to encode the same protein product raises the question of whether the 2.7 kb blastoderm-specific RNA plays a role distinct from that of the 4.5 kb RNA. We have begun experiments to test whether the two RNAs are distributed differently in the embryo.

The hydrophobicity correlations suggest that there is a structural relationship between the similar domains of the bsg25D and fos proteins. While the similarity of these domains may arise, in part, from their predicted extensive alpha-helical structure (20), it may also indicate that the bsg25D protein has a function related to that of the fos protein. Arguments suggesting that the similarities arise not only from alpha-helical structure are that: 1) of all the alpha-helical proteins present in the database, none were nearly as similar as the two discussed here, and 2) other regions of the bsg25D amino acid sequence which are predicted to be equally as alpha-helical as the fos and tropomyosin similarity domains are not similar to these proteins. In any case, it is interesting to speculate briefly about the implications that these similarities, if they represent a functional relationship, might have for the developmental role of the bsg25D locus.

The similarity of a small bsg25D protein segment to repeated segments of tropomyosin which are thought to bind actin raises the possiblity that the bsg25D protein might have actin-binding properties. This could be important during the blastoderm stage when dramatic cytoskeletal reorganizations, including cell formation, are occurring. It has been suggested that actin-binding domains might function in early embryogenesis to localize molecular determinants in the embryo (33).

The fos gene is a member of the competence gene family--genes induced by platelet-derived growth factor. The fos protein is present in the nucleus and has been speculated to play a role in in signalling cells to cease dividing prior to differentiation (34,35). The domain of the fos oncogene product which we show here to be similar to the bsg25D gene product has also been shown to be homologous to several other members of the competence gene family, including r-fos and the myc homologs (36,37, Fig. 5). Although the bsg25D, r-fos, and myc homologs do not share amino acid sequence homology, hydrophobicity correlation coefficient analysis suggests that the bsg25D and c- and v-myc domains shown in Fig. 5 do share similar three-dimensional structures (20). The structural similarity to products of several genes involved in changes in the cell cycle and in differentiation may be relevant to the function of the bsg25D gene, which is expressed during a period of embryogenesis when the rate of nuclear division is slowing dramatically and cell commitment is taking place.

The data presented here form the basis for further investigations into the function of the bsg25D locus. Studies are underway to characterize the spatial and temporal localization of the bsg25D RNAs, and antisera have been raised against a peptide predicted from the sequence data (20) for use in similar characterizations of the protein in developing embryos. A recent genetic analysis of chromosomal region 25A-F (J. Szidonya and G. Reuter, personal communication) provides mutations which can be tested for relationship to the bsg25D locus. The addition of new sequence information to protein databases may reveal further sequence relationships. We expect that results from these different experimental approaches will provide clues about the function of the bsg25D protein and its role in embryogenesis.

ACKNOWLEDGMENTS

We thank Debra Peattie for assistance with sequencing technology, Ken Burtis for a codon frequency table, Russell Doolittle for useful discussions, and Richard Baldarelli, Kathryn Calame, and David Eisenberg for useful suggestions during the progress of this work and on the manuscript. Some sequence analysis was done using BIONET (USPHS RR01685). This work was supported by USPHS Grants GM07185 to P.D.B. and P.A.M., HD09948 to J.A.L., and an NSF graduate fellowship to P.A.M.

*Present address: Division of Biology, California Institute of Technology, Pasadena, CA 91125, USA

+To whom correspondence should be addressed

REFERENCES

1. Lengyel, J., Roark, M., Kongsuwan, K., Mahoney, P., Boyer, P. and Merriam, J. (1985) In Sawyer, R. and Showman, R. (eds), The Cellular and Molecular Biology of Invertebrate Development, U.S.C. Press, Columbia, pp. 239-258.
2. Roark, M., Mahoney, P., Graham, M. and Lengyel, J. (1985) Dev. Biol. 109, 476-488.
3. Vincent, A., Colot, H. and Rosbash, M. (1985) J. Mol. Biol. 186, 149-166.
4. Vincent, A. (1986) Nucl. Acid. Res. 14, 4385-4391.
5. Maniatis, T., Fritsch, E. and Sambrook, J. (1982) Molecular Cloning. A Laboratory Manual. Cold Spring Harbor Press, Cold Spring Harbor.
6. Sanger, F., Coulson, A., Barrell, B., Smith, A. and Roe, B. (1980) J. Mol. Biol. 143, 161-178.
7. Biggin, M., Gibson, T. and Hong, G. (1983) Proc. Natl. Acad. Sci. USA 80, 3963-3965.
8. Yanisch-Perron, C., Vieira, J. and Messing, J. (1985) Gene 33, 103-119.
9. Hong, G. (1982) J. Mol. Biol. 158, 539-549.
10. Henikoff, S. (1984) Gene 28, 351-359.
11. Deininger, P. (1983) Anal. Biochem. 129, 216-223.
12. Poole, S., Kauvar, L., Drees, B. and Kornberg, T. (1985) Cell 40, 37-43.
13. Benton, W. and Davis, R. (1977) Science 196, 180-182.
14. Boyer, P. (1986) Nucl. Acid. Res. 14, 7505.

15. Hultmark, D., Klemenz, R. and Gehring, W. (1986) Cell **44**, 429-438.
16. Berk, A. and Sharp, P. (1977) Cell 12, 721-732.
17. Staden, R. (1980) Nucl. Acid. Res. **8**, 3673-3694.
18. Staden, R. (1982) Nucl. Acid. Res. **10**, 4731-4751.
19. Staden, R. (1984) Nucl. Acid. Res. 12, 521-538.
20. Boyer, P. (1986) Ph.D. Thesis, University of California, Los Angeles.
21. Wilbur, W. and Lipman, D. (1983) Proc. Natl. Acad. Sci. USA **80**, 726-730.
22. George, D., Barker, W. and Hunt, L. (1986) Nucl. Acid. Res. **14**, 11-15.
23. Orcutt, B., Dayhoff, M., George, D. and Barker, W. (1984) PIR Report ALI-1284.
24. Kabsch, W. and Sander, C. (1984) Proc. Natl. Acad. Sci. USA **81**, 1075-1078.
25. Doolittle, R. (1981) Science **214**, 149-159.
26. Sweet, R. and Eisenberg, D. (1983) J. Mol. Biol. **171**, 479-488.
27. Mount, S. (1982) Nucl. Acid. Res. **10**, 459-472.
28. Corden, J., Wasylyk, B., Buchwalder, A., Sassone-Corsi, P., Kedinger, C. and Chambon, P. (1980) Science **209**, 1406-1414.
29. Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. and Chambon, P. (1978) Proc. Natl. Acad. Sci. USA **75**, 4853-4857.
30. Sheterline, P. (1983) Mechanisms of Cell Motility: Molecular Aspects of Contractility, Academic Press, San Francisco.
31. Simonsen, C. and Levinson, A. (1983) Mol. Cell. Biol. **3**, 2250-2258.
32. Mason, P., Jones, M., Elkington, J. and Williams, J. (1985) EMBO J. **4**, 205-211.
33. Miller, K., Karr, T., Kellogg, D., Mohr, I., Walter, M. and Alberts, B. (1985) C.S.H. Symp. Quant. Biol. **50**, 79-90.
34. Muller, R., Bravo, R., Burckhardt, J. and Curran, T. (1984) Nature **312**, 716-720.
35. Mitchell, R., Henning-Chubb, C., Huberman, E. and Verma, I. (1986) Cell **45**, 497-504.
36. Cochran, B., Zullo, J., Verma, I. and Stiles, C. (1984) Science **226**, 1080-1082.
37. Kohl, N., Legouy, E., DePinho, R., Nisen, P., Smith, R., Gee, C. and Alt, F. (1986) Nature **319**, 73-77.
38. Dayhoff, M. (1979) Atlas of Protein Sequence and Structure, Vol. 5, Suppl. 3.
39. Wolfenden, R., Cullis, P. and Southgate, C. (1979) Science **206**, 575-577.
40. Wolfenden, R., Andersson, L., Cullis, P. and Southgate, C. (1981) Biochemistry **20**, 849-855.
41. Janin, J. (1979) Nature **277**, 491-492.