Proc. Natl. Acad. Sci. USA
Vol. 83, pp. 1827–1831, March 1986
Genetics

Structure and expression of the murine N-*myc* gene

(genomic sequence/transient expression/c-*myc* protooncogene)

RONALD A. DEPINHO, EDITH LEGOUY, LORI B. FELDMAN, NANCY E. KOHL, GEORGE D. YANCOPOULOS, AND FREDERICK W. ALT

Department of Biochemistry and Molecular Biophysics, Columbia University, College of Physicians and Surgeons, New York, NY 10032

Communicated by Matthew D. Scharff, November 14, 1985

ABSTRACT **We have demonstrated that the entire murine N-*myc* gene and the sequences necessary for its expression in human neuroblastoma cells are contained within a 7.4-kilobase murine genomic clone. The complete nucleotide sequence of this gene reveals a number of striking similarities and differences when compared to the related c-*myc* gene including the following: (*i*) each gene contains three exons of which the first encodes a long 5′-untranslated leader sequence; (*ii*) the coding regions of the N- and c-*myc* genes share regions of substantial nucleic acid homology, the putative N-myc protein shares substantial homology with the c-myc protein; (*iii*) as with c-*myc*, extensive nucleotide sequence homology exists between the untranslated regions of the human and murine N-*myc* gene transcripts; however, the N-*myc* and c-*myc* untranslated regions are totally divergent; (*iv*) the N-*myc* transcriptional promoter differs from that of c-*myc* and is more related to the promoter of the simian virus 40. We discuss these findings in the context of previously defined similarities and differences in the potential functional and regulatory aspects of these two *myc*-family members.**

N-*myc* is a cellular gene that has been implicated in the development of a highly restricted set of tumors, most notably, neuroblastoma and retinoblastoma. A causal role for the N-*myc* gene in carcinogenesis is suggested by its frequent amplification or overexpression in many of these tumors (1–5). Functional and structural evidence suggests that the N-*myc* and c-*myc* genes may be members of a larger *myc* oncogene family. In particular, N-*myc* bears significant sequence homology to c-*myc* (1, 6) and has comparable transforming potential in the rat embryo fibroblast assay (7, 8). Despite these similarities, N-*myc* expression appears only in a restricted set of tissues and tumors, while c-*myc* expression is more generalized, suggesting that the two *myc* genes have different physiological roles (3, 29). We have previously described a murine genomic clone that appeared to encode a complete and functional N-*myc* gene product when incorporated into a retrovirus expression vector (7). To further define the relationship between the N-*myc* and c-*myc* genes, we now have determined the complete nucleic acid sequence of this clone. In addition, we clearly demonstrate by transfection studies that this clone contains not only the entire N-*myc* gene but also the flanking sequences necessary to support its expression in neuroblastoma cell lines. Our findings demonstrate that the N-*myc* and c-*myc* genes and gene products have a similar general organization but have major divergences within potential regulatory regions.

MATERIALS AND METHODS

Cell Culture. Human neuroblastoma lines, LAN-5, LAN-1, and SKNSH, and the mouse Abelson-transformed pre-B cell line, 38B9, were maintained as described (2, 9).

Eukaryotic Cell Transfection. On the day prior to transfection, cultures were seeded at 2×10^6 cells/10 ml of media. Four hours before transfection, media was switched from RPMI/12% (vol/vol) fetal calf serum to Dulbecco's modified Eagle's media/12% (vol/vol) fetal calf serum. Transfections were carried out by using the calcium phosphate precipitation technique as described (7), except that carrier DNA was not added and 10 μg of plasmid DNA was used per plate. Following transfection, neuroblastoma cell lines were maintained in RPMI/12% (vol/vol) fetal calf serum. At 48 hr after transfection, cells were harvested and total RNA was extracted as described (7). The genomic clone pN7.7 with and without an associated Moloney leukemia virus enhancer (long terminal repeat, LTR) was described (7).

Nucleic Acid Preparation and Nucleotide Sequencing. Plasmid preparation, restriction endonuclease digestion, and mapping were as described (7). Sequencing was performed by the method of Maxam and Gilbert (10).

S1 Nuclease Assays and Probes. The probes are shown in Fig. 1. Single-stranded uniformly labeled DNA probes were prepared, hybridized, digested, and analyzed by electrophoresis through a 5% acrylamide/7 M urea gel as described (8, 11).

Computer Analysis of DNA and Protein Sequences. Graphic matrix analysis was performed with the MBSP dot matrix program written at the Albert Einstein College of Medicine. Percent homologies were calculated according to an alignment program as described (12).

RESULTS AND DISCUSSION

The Mouse N-*myc* Gene Contains Three Exons. We have suggested that the murine genomic clone pN7.7 contains a complete copy of the N-*myc* gene (7). Comparison of the nucleotide sequence of this clone (Fig. 1, *Upper*) to that of the human N-myc cDNA (8) supports this suggestion and indicates that the N-*myc* gene is organized into three exons and two introns (Fig. 1, *Lower*). This proposed organization is based on several lines of evidence. A dot matrix computer analysis comparing the human cDNA sequence and the complete murine N-*myc* gene sequence indicates that the mouse gene contains three regions of significant homology with human cDNA sequence and that each region of homologous sequence is separated by large stretches of unrelated sequence (Fig. 2*A*). At each of the boundaries between conserved (putative exons) and divergent sequence (putative introns), there exists a typical donor or acceptor splice recognition sequence for eukaryotic genes (13). The 5′ boundaries of the first and third exons were defined precisely through S1 nuclease mapping analyses that utilized a probe spanning the putative 5′ boundary of the first exon and a probe spanning the putative 5′ boundary of the third exon; the sizes of the fragments [185 base pairs (bp) and 553 bp,

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

Abbreviations: bp, base pair(s); kb, kilobase(s); LTR, long terminal repeat(s).

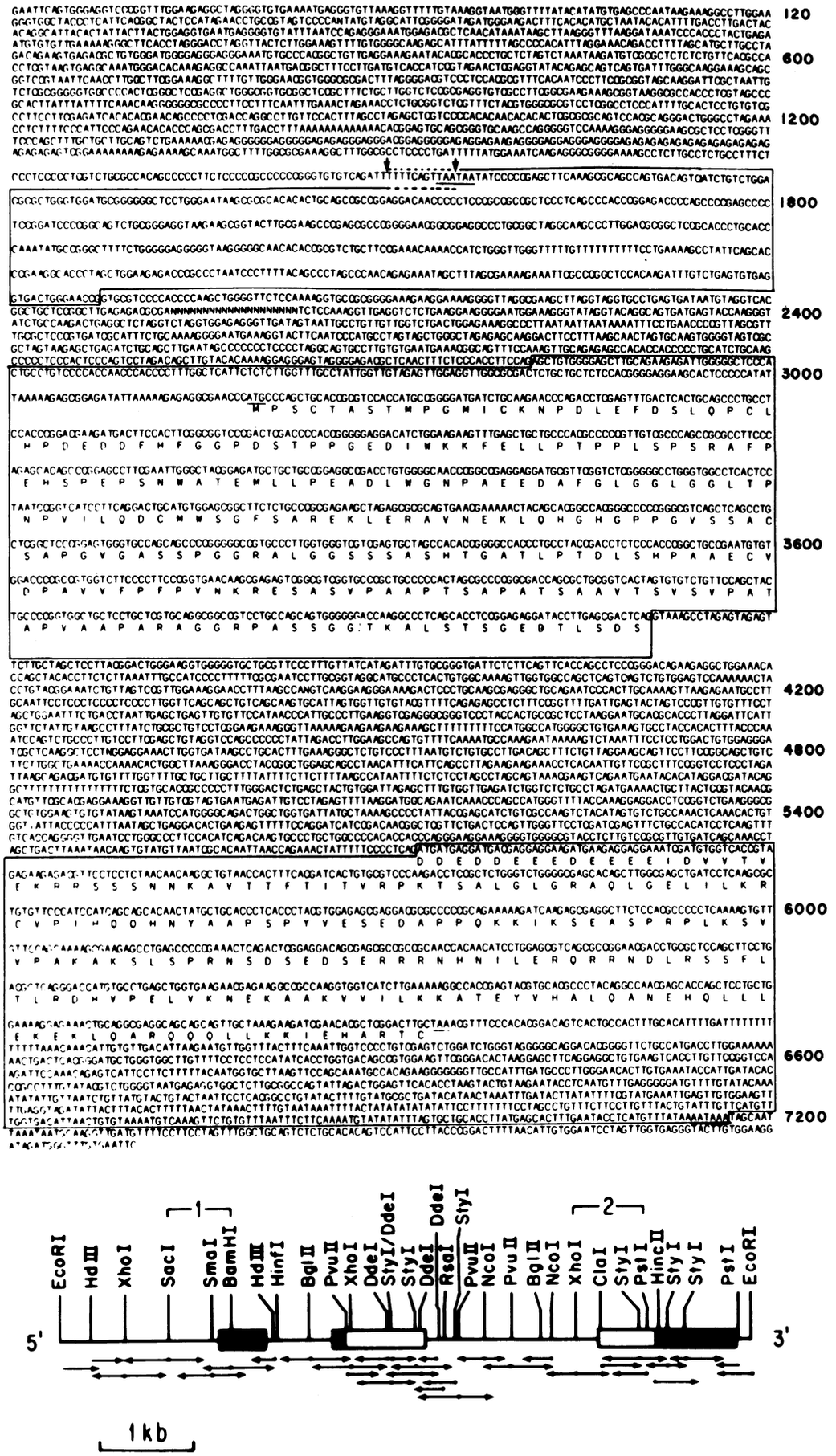

FIG. 1. (*Legend appears at the bottom of the opposite page.*)

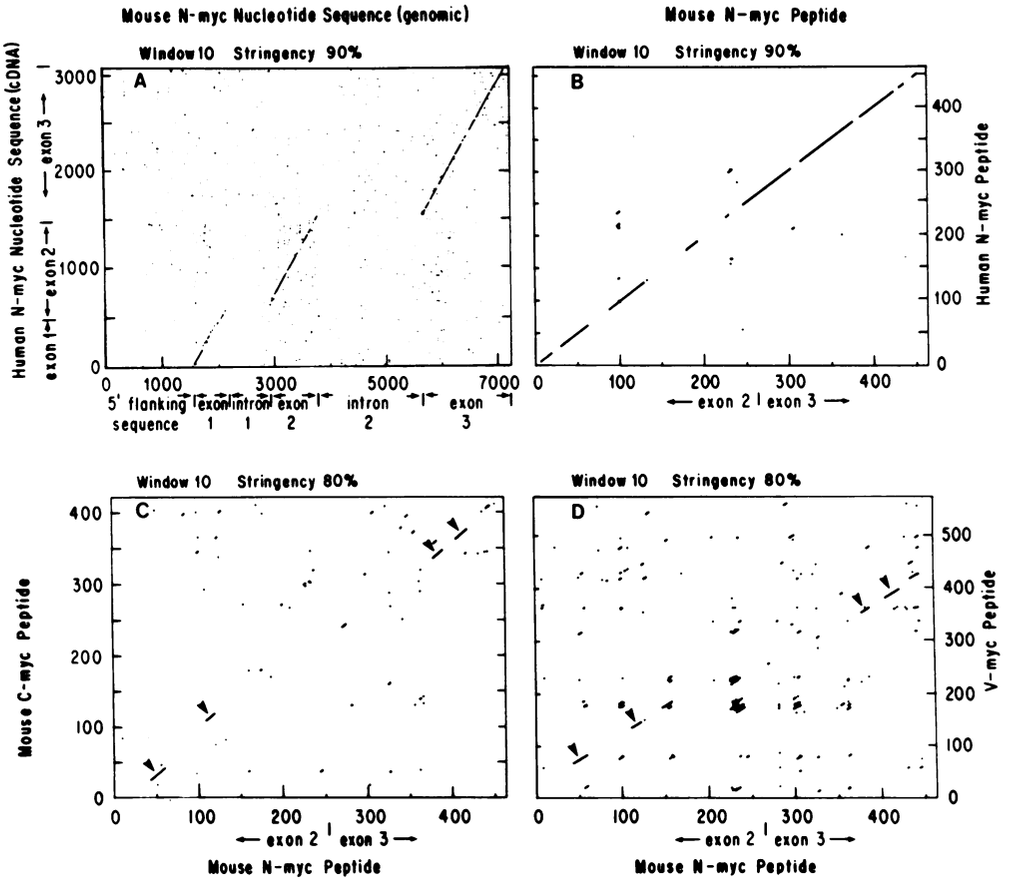

FIG. 2. Dot matrix computer analysis of nucleotide and peptide sequence homology among members of the *myc* oncogene family. (*A*) Nucleotide sequence comparison between human N-myc cDNA and mouse N-myc genomic clones. Peptide comparison between human and mouse N-myc (*B*), mouse N-myc and mouse c-myc (*C*), mouse N-myc and v-myc (*D*). Comparisons in all panels were conducted at a stringency of 80% or 90% with a window of 10 residues where a match of 8 of 10 or 9 of 10 residues, respectively, is required for placement of a dot. Lines of homology represent a continuous series of dots that correspond to regions in the compared sequences that are 80 to 90% homologous over a stretch of 10 residues.

respectively] protected by RNA from an N-*myc*-expressing pre-B cell firmly established the 5′ location of exon 3 at position 5688 and that of exon 1 at about position 1625 (Fig. 3; see below for further description of promoter region). Additional S1 nuclease assays and RNA blotting analyses indicated that no additional coding regions exist to the 5′ side of the putative upstream boundary of exon 1 (data not shown). At the 3′ end of the gene, the location of the adenylylation signal was determined by a sequence comparison between the mouse genomic sequence and the corresponding 3′ terminus of human cDNA sequence (8). A consensus adenylylation sequence, AATAAA (14), occurs at the same location in mouse and human genes (positions 7188 through 7193). Thus, the entire mouse N-*myc* gene spans ≈5569 bp from cap site(s) to polyadenylylation signal and consists of three exons (exon 1, ≈550 bp; exon 2, 979 bp; and exon 3, 1489 bp) separated by two introns (first intron, ≈670 bp; second intron, 1883 bp); the length of the predicted message of ≈3.0 kilobases (kb) agrees well with the published size of the mature murine N-myc message (7).

Expression of the N-*myc* Gene: Promoter and 5′-Flanking Region. To define the sequences necessary for the expression of the N-*myc* gene, we have transfected the pN7.7 clone with and without an associated LTR into various N-*myc*-expressing human neuroblastoma cell lines. Transient expression was assayed via the S1 nuclease method outlined above. RNA isolated from both the pN7.7- and pN7.7+LTR-transformed, but not mock-transformed, human neuroblastomas protected a portion of the 5′-exon 1 and 5′-exon 3 probes; the protected fragment was identical in size to that protected by authentic N-myc mRNA from the N-*myc*-expressing pre-B cell line (Fig. 3). Given that the probes were uniformly labeled, the correct location of the protected fragment was deduced through sequence comparison with the human N-myc cDNA (8). These results demonstrate that the pN7.7 clone contains the necessary sequences to direct the correctly initiated expression of the N-*myc* gene in neuroblastomas and also demonstrate that N-*myc* transcription is initiated from the same promoter in pN7.7+LTR transformants. Furthermore, in preliminary experiments, we were unable to detect N-*myc* expression in the 3T3 murine fibroblast cell line unless the pN7.7 clone was associated with an LTR (data not shown). Our current findings are in agreement with the previous suggestion that an associated

FIG. 1. Organization and nucleotide sequence of the mouse N-*myc* gene. The sites of transcriptional initiation are indicated by the dashed portion of box 1. Adenylylation signal and start codon are underlined. The single-letter amino acid sequence code is shown below the mouse nucleic acid sequence. The sequencing strategy is shown below the partial restriction map that depicts the overall organization of the pN7.7 clone; darkened areas indicate untranslated regions. Coding domains (open areas) were sequenced in both directions. S1 probes (probe 1 = 0.66-kb *Sac* I–*Bam*HI; probe 2 = 0.70-kb *Xho* I–*Pst* I) used in transient expression studies and S1 nuclease mapping assays are indicated above the restriction map.

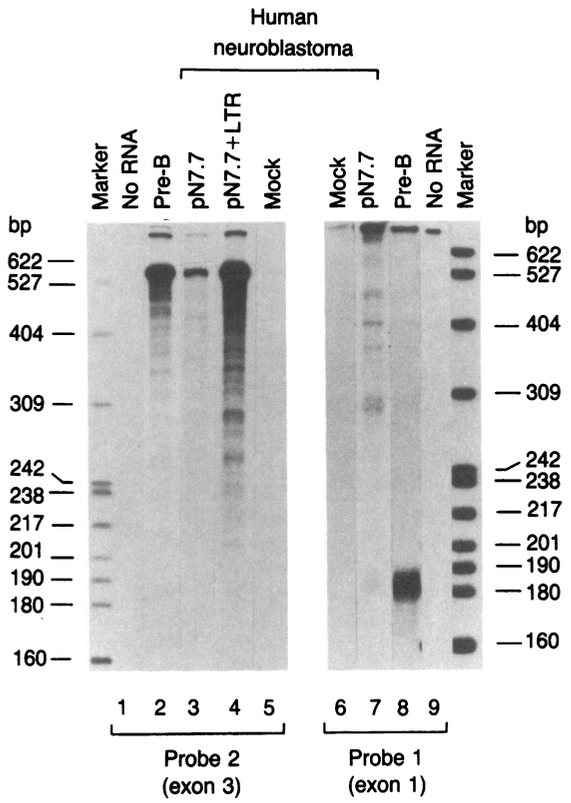

FIG. 3. Expression of the N-*myc* gene. The murine genomic clone pN7.7, with and without a retroviral LTR, was transfected into various human neuroblastoma cell lines; one representative experiment is shown, using the LAN-5 line. The no RNA lane represents a negative control for the S1 mapping analysis; "Mock" indicates RNA prepared from cells that did not receive transfected DNA; pre-B is RNA prepared from pre-B cell line, 38B9; pN7.7 and pN7.7+LTR represents RNA prepared from lines transfected with the respective plasmids. (Lane 7 represents a longer exposure making artifactual bands in the >300 bp range more evident; these bands were identical in the other lanes on longer exposure.)

retroviral enhancer region was necessary to support expression of the pN7.7 clone in rat embryo fibroblasts (7) and further suggest that specific regulatory elements contained with the pN7.7 sequence may mediate the tissue- and stage-specific expression of the N-*myc* gene.

The reproducibly diffuse band protected when the 5′-exon 1 probe was employed in the S1 analyses indicated that there are several transcription initiation sites within a 10-bp region encompassing position 1625 (Fig. 3). Both primer extension and S1 nuclease mapping analyses have mapped transcription initiation of the human N-*myc* gene to this region, but in the case of the human gene as many as 10–15 distinct initiation sites were found (8). The c-*myc* gene appears to contain two independent promoter regions that contain CAAT and TATA boxes upstream from a single initiation site (15). The TATA box is believed to be important in determining the site of transcriptional initiation (16); unlike c-*myc*, the promoter region of the murine N-*myc* gene does not appear to contain upstream TATA or CAAT motifs, although an A+T-rich (TATA-like) region exists just downstream to the cap site. Significantly, inspection of the nucleotide sequence in the vicinity of and just upstream from the N-myc transcription initiation sites revealed a region, starting at the A+T-rich region and extending 115 bases upstream, which has nearly 60% nucleotide sequence homology to the simian virus 40 transcriptional promoter (Fig. 4) (16); this 5′ region contains copies of the simian virus 40 promoter hexanucleotide repeat CCGCCC and its complement GGGCGG. Simian virus 40-like promoters have been identified in a number of other cellular genes (17–22). Of particular interest in this region is a 14-bp palindromic sequence that contains the CCGCCC/GGGCGG elements. These sequences occur in the murine N-*myc* and c-*myc* promoter regions in analogous positions (Fig. 4); the occurrence of these conserved sequences is particularly striking in the context of the very divergent nature of the remainder of the 5′ N-*myc* and c-*myc* sequences and the sequences that encode the untranslated regions of the two messages. Approximately 150 bp upstream from the N-*myc* transcription initiation region is a repetitive ≈120-bp region consisting of the core unit, (A-G)*n* (where *n* = 1 to 6). Although the role of this 5′-repeat sequence, if any, is unknown, it is notable that a repetitive nonanucleotide (C/TCC/TCCCCT) occurs in the same 5′ location in the c-*myc* gene (15).

Structure of the N-myc Transcript. The first exon of the N-*myc* gene contains several potential translation initiation codons, but all are followed within exon 1 by inphase translation termination codons. The first ATG sequence of a long open reading frame occurs approximately 180 bp downstream from the 5′ boundary of exon 2 (position 3038); beginning at this position the open reading frame extends 1386 nucleotides to an inphase terminator at position 6426 in exon 3. The location of this potential coding domain corresponds precisely to the coding region defined in the human N-myc cDNA (8). We have assigned the first ATG as coding for the N-terminal methionine residue of the protein although it is possible that initiation can occur at two other ATG codons that lie inphase at positions 3062 and 3071. Thus, the mature 3.0-kb mouse N-myc mRNA consists of (*i*) an unusually large 5′-untranslated leader (750 bp) encoded by the first exon and the 5′ portion of exon 2; (*ii*) a 1.4-kb coding domain spanning exons 2 and 3; and (*iii*) a large 3′-untrans-

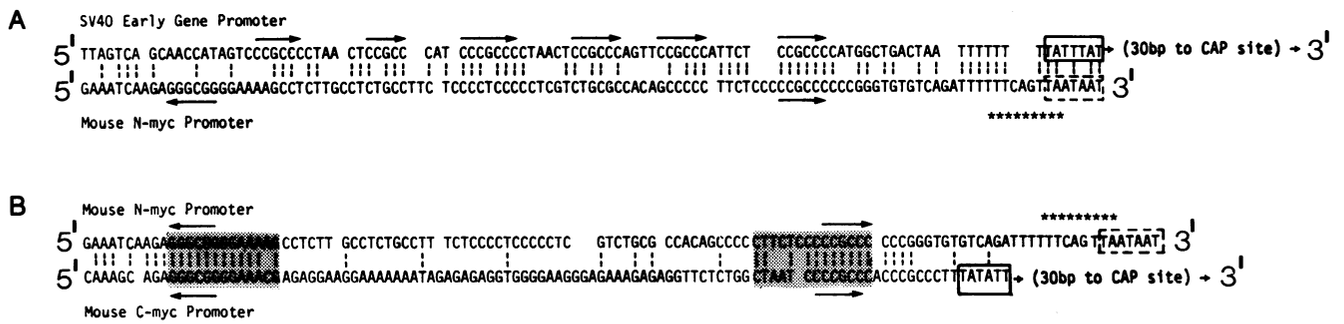

FIG. 4. Sequence homology between the promoter region of (*A*) N-*myc* versus the early promoter region of simian virus 40 (SV40) and (*B*) N-*myc* versus c-*myc* promoters. Alignment of these sequences was arranged to achieve maximum homology. The hexanucleotides (CCGCCC) and its complement (GGGCGG) are indicated by arrows. The shaded regions of homology in the N-*myc* and c-*myc* promoters are near inverted repeats of each other. ★, cap sites.

lated region (900 bp) encoded by the downstream portion of exon 3. The presence of large untranslated regions upstream from the coding domain is not unprecedented in other eukaryotic genes; most notably, the first exon of the c-*myc* gene also encodes a large 5'-untranslated leader sequence (15, 23).

A striking feature of both 5'- and 3'-untranslated regions of the N-*myc* gene is the high degree of homology seen between mouse and human genes. If allowance is made for insertions and deletions, the overall nucleotide homology between N-*myc* genes is ≈80% for the 5' leader and ≈90% for the 3'-untranslated region, values that are remarkably high considering that the exon 2 coding region is less conserved (see below). Significant homology is also encountered in the mouse and human c-*myc* untranslated sequences (24). The high degree of nucleotide sequence conservation in regions that are not under any selective pressure at the protein level suggests that they may possess some important regulatory role in the expression of the *myc* genes. However, although there is significant conservation of N-*myc* and c-*myc* coding regions (see below), there is no significant homology between the untranslated regions. Thus, if the conserved 5'- and 3'-untranslated regions of the N-*myc* and c-*myc* genes are involved in regulatory processes, the divergence between these sequences may play a role in the dramatically different expression patterns of the two genes (29).

The N- and c-myc Proteins Share Significant Homology. The putative N-myc transcript could encode a protein of 462 amino acids with a predicted molecular size of approximately 50 kDa. If allowance is made for insertions and deletions, the overall homology between the mouse and human proteins is greater than 85%. The overall amino acid homology of the mouse N-myc to mouse c-myc (23) and v-myc (25) is 35% and 34%, respectively. In particular, a dot matrix computer analysis reveals clusters of amino acids that are highly conserved across all myc proteins (arrows in Fig. 2 *C* and *D*) including two that were previously noted in the 5' portion of the coding domain in exon 2 (28) and two additional ones in the distal coding domain of exon 3. The human and murine c-*myc* gene products have been shown to be nuclear-associated and contain DNA binding capacity *in vitro* (26). This property has been attributed to an abundance of basic amino acids at the carboxyl terminus that could account for an affiliation of the protein product with chromatin (27). An examination of the hydrophilicity of the putative N-*myc* gene product reveals a similar pattern. It is, therefore, possible that the N-myc protein may also be a DNA-binding protein.

The *myc* Gene Family. There are a number of similarities with respect to structure and function of the N- and c-*myc* genes and gene products. The gene products have a similar transforming activity and share several highly conserved regions, at least one of which was implicated in the DNA binding activity of c-myc. The genes also have a similar organization; both have three exons of which the first encodes a long and highly conserved untranslated leader sequence. However, despite these remarkable similarities in general structure and function, the genes show striking differences with respect to the regulation of their expression. The N- and c-*myc* genes belong to a larger *myc* gene family that also includes the L-*myc* gene (28) and probably many others; these genes are differentially expressed with respect to tissue and developmental stage with N- and L-*myc* expression being much more restricted than that of c-*myc* (29). Our current studies support the possibility that restricted expression of the N-*myc* gene may be mediated by sequences contained within or near the gene. In contrast to the similarities in structure and function of the N- and c-*myc* genes, the divergent nature of the potential regulatory portions of these genes may provide a basis for their divergent patterns of expression.

This work was supported by National Institutes of Health Grant 2-POI CA 23767-06, by the American Cancer Society Grant CD-269, and a Searle Scholars Award (F.W.A.). R.A.D. is a recipient of the Physician–Scientist Award National Institutes of Health AI00602. F.W.A. is a Mallinkrodt Scholar. E.L. is a European Molecular Biology Organization fellow. We are extremely grateful to Kenneth Krauter for his assistance with the computer analysis.

1. Schwab, M., Alitalo, K., Klempnauer, K. H., Varmus, H., Bishop, J. M., Golbert, F., Brodeur, G., Goldstein, M. & Trent, J. (1983) *Nature (London)* **305,** 245–248.
2. Kohl, N. E., Kanda, N., Schreck, R. R., Bruns, G., Latt, S. A., Gilbert, F. & Alt, F. W. (1983) *Cell* **35,** 359–367.
3. Kohl, N. E., Gee, C. E. & Alt, F. W. (1984) *Science* **226,** 1335–1337.
4. Brodeur, G. M., Seeger, R. C., Schwab, M., Varmus, H. E. & Bishop, J. M. (1984) *Science* **224,** 1121–1124.
5. Lee, W. H., Murphee, A. L. & Benedict, W. F. (1984) *Nature (London)* **309,** 458–460.
6. Michitsch, R. W. & Melera, P. W. (1985) *Nucleic Acids Res.* **13,** 2545–2558.
7. Yancopoulos, G. D., Nisen, P. D., Tesfaye, A., Kohl, N. E., Goldfarb, M. P. & Alt, F. W. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 5455–5459.
8. Kohl, N. E., Legouy, E., DePinho, R. A., Nisen, P. D., Smith, R. K., Gee, C. E. & Alt, F. W. (1986) *Nature (London)* **319,** 73–77.
9. Alt, F., Rosenberg, N., Lewis, S., Thomas, E. & Baltimore, D. (1981) *Cell* **27,** 381–390.
10. Maxam, A. and Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 560–564.
11. Biggin, M., Farell, P. J. & Barrell, B. G. (1984) *EMBO J.* **3,** 1083–1090.
12. Wilbur, W. J. & Lippman, D. J. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 726–730.
13. Mount, S. M. (1982) *Nucleic Acids Res.* **10,** 459–472.
14. Proudfoot, N. J. & Brownlee, G. G. (1976) *Nature (London)* **263,** 211–214.
15. Battey, J., Moulding, C., Taub, R., Murphy, W., Stewart, T., Potter, H., Lenoir, G. & Leder, P. (1983) *Cell* **34,** 779–787.
16. Mathis, D. & Chambon, P. (1981) *Nature (London)* **290,** 310–315.
17. Osborne, T. F., Goldstein, J. L. & Brown, M. S. (1985) *Cell* **42,** 203–212.
18. Valerio, D., Duyvesteyh, M. G. C., Dekker, B. M. M., Weeda, G., Berkvens, Th. M., Van der Voorn, L., Van Ormondt, H. & Van der Eb, A. J. (1985) *EMBO J.* **4,** 437–443.
19. McGrogan, M., Simonsen, C. C., Smouse, D. T., Farnham, P. J. & Schimke, R. T. (1985) *J. Biol. Chem.* **260,** 2307–2314.
20. Levanon, D., Lieman-Hurwitz, J., Dafni, N., Wigderson, M., Sherman, L., Berstein, Y., Laver-Rudich, Z., Danciger, E., Stein, O. & Groner, Y. (1985) *EMBO J.* **4,** 77–84.
21. Roebuck, K. A. & Stump, W. E. (1985) *DNA* **4,** 86.
22. Melton, D. W., Konecki, D. S., Brennard, J. & Caskey, C. T. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 2147–2151.
23. Stanton, L. W., Farlander, P. D., Tesser, P. M. & Marcu, K. B. (1984) *Nature (London)* **310,** 423–425.
24. Bernard, O., Cory, S., Gerondakis, S., Webb, E. & Adams, J. M. (1983) *EMBO J.* **2,** 2375–2383.
25. Alitalo, K., Bishop, J. M., Smith, D. H., Chen, E. Y., Colby, W. W. & Levinson, A. D. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 100–104.
26. Persson, H. & Leder, P. (1984) *Science* **225,** 718–721.
27. Alitalo, K., Ramsay, G., Bishop, J. M., Pfeifer, S. O., Colby, W. W. & Levinson, A. D. (1983) *Nature (London)* **306,** 274–277.
28. Nau, M. M., Brooks, B. J., Battey, J., Sausville, E., Gazdar, A. F., Kirsch, I. R., McBride, O. W., Bertness, V., Hollis, G. F. & Minna, J. D. (1985) *Nature (London)* **318,** 69–73.
29. Zimmerman, K. A., Yancopoulos, G. D., Collum, R. G., Smith, R. K., Kohl, N. E., Denis, K. A., Nau, M. M., Witte, O. N., Toran-Allerand, D., Gee, C. E., Minna, J. D. & Alt, F. W. (1986) *Nature (London)*, in press.