Vol. 140, No. 1, 1986 BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS
October 15, 1986 Pages 313-319

ON THE HIGH CONSERVATION OF THE HUMAN c-myc FIRST EXON

Marthe E. ELADARI, SAEED H. SYED *, Stephane GUILHOT, Luc d'AURIOL and Francis GALIBERT

Laboratoire d'Hématologie Expérimentale, Centre Hayem, Hôpital Saint-Louis, 2, place du Dr. Fournier, 75475 PARIS Cédex 10, France

Received August 18, 1986

Several hypotheses have been proposed to explain the role that the human c-myc 1st exon is thought to play in the regulation of the expression of the c-myc protein. One of these hypotheses, based on sequencing data showing an open reading frame overlapping that exon, predicts the existence of a protein most probably involved in the regulation of the expression of the c-myc protein. However, mainly because several other published sequences are devoided of such an ORF, this hypothesis did not retain much attention. In this paper, we present two additional sequences fully identical to the 1st exon region sequence we previously published, and discuss the implication of such a high degree of conservation for the human c-myc 1st exon.

The cellular myc oncogene is expressed in a variety of tissues and, like several other oncogenes, the c-myc gene is thought to play an important role in the control of proliferation of normal cells (1). Using a centrifugal elutriation technique, Hann et al. (2) were able to show that the synthesis, half life and modification of c-myc protein are constant throughout the cell cycle of normal and transformed cell. On the other hand, mitogenic stimulation of quiescent lymphoid, fibroblast or epithelial cells apparently leads to a sharp increase in c-myc RNA levels (3, 4, 5). Taken together, these data suggest that a burst of c-myc expression occurs during the initiation of growth and that subsequently a constant basal level is maintained during exponential growth.

The normal human (6), mouse (7) and chicken (8) myc genes are made of three exons. Exons 2 and 3 code for the c-myc protein, a nuclear

* Present address : Cancer Studies Section, King Faisal Specialist Hospital and Research Center 11211 RIYAD SAUDI ARABIA.

0006-291X/86 $1.50

phosphoprotein (9), homologous to the v-myc protein coded by the transforming sequence of avian myelocytomatosis virus (10).

Two promotors, P1 and P2, about 150 base pairs apart, are used with different frequencies in normal and transformed cells to initiate transcription of the c-myc gene (11). Starting from P1, there is a long leader of 550 bp corresponding to the 1st exon (11, 12, 13). Several observations have suggested the implication of the 1st exon in the regulation of c-myc expression (14, 15). In 75 % of Burkitt lymphoma cell lines a translocation occurs between chromosomes 8 and 14 with a breakpoint located on chromosome 8 in the vicinity of exon 1 (16). As a result of this translocation exon 1 is either separated from the remaining part of the gene or shows nucleotide sequence modifications (15,17,18,19). In the remaining cases, translocations were observed implicating either chromosomes 2 or 22 with again numerous mutations occuring within exon 1. The concept of transcriptional deregulation by alteration of exon 1 (20, 21) was also supported by the finding that in most Burkitt's lymphoma lines the c-myc gene on the normal chromosome remained silent while the altered c-myc gene was transcribed (14, 15, 22, 23).

Several mechanisms have been proposed to explain the role of this leader. It has been suggested that a repressor molecule, might bind to the 1st exon (15, 20, 21, 24) or that the 1st exon might be involved, through different secondary structures, in the control of protein synthesis (25). A third possibility has recently emerged from the finding by Rabbitts et al (26) and Eick et al. (27) that truncation of exon 1 from the c-myc gene results in a longer half life for c-myc RNA.

In a previous work (28) we have noted the existence of a large ORF overlapping the 1st exon, and have proposed that this ORF could be used by the cell to make a protein of 188 amino acids. As suggested by the close relationship of this ORF with the c-myc locus, this protein might play a role connected to the regulation of the c-myc protein.

However, this last hypothesis has not retained much attention, mainly because other sequence data have reported the existence of stop

codons in the three reading frames of this exon (11, 13, 15, 22). This discrepancy in the sequence data might be due to sequence polymorphism, to cloning or sequencing problems or to the choice of the tissue of origin. In order to further check that hypothesis, we have cloned and sequenced the two alleles of the lymphocytes of a healthy adult donor. We report here these sequences and compare them with the sequences already published.

MATERIAL AND METHODS

Lymphocytes from a healthy adult of caucasian origin were kindly obtained through the Centre d'Etude du Polymorphisme Humain (CEPH). High molecular DNA was extracted by the sarcosyl-proteinase K lysis method followed by phenol and RNase treatments (29). Extracted DNA was partially digested with Sau3a and fractionated by sucrose gradient. DNA fragments in the size range of 15 to 20 Kb were recovered and ligated overnight to EMBL4 arms. Ligated DNA was in vitro packaged with λ protein extract (Amersham) and used to infect E. coli. 2.5×10^6 independent recombinants were obtained and amplified. Screening of the library with the human 1st exon cloned in M13 was performed following conventional methods (29).

5 λ c-myc recombinants were isolated, plaque purified and used to make stocks. DNAs were prepared from all 5 clones and used for restriction enzyme mapping and sequencing. Fragments corresponding to the 1st exon were subcloned in M13mp8 in both orientations. 8 specific primers 20 to 25 mers, 4 for each strand, regularly spaced on the 1st exon myc sequence, were synthesized by the phosphotriester method (30) using a SAM1 (Biosearch) automate. These oligonucleotides were used along with the universal primer to derive the nucleotide sequence of the 1st exon on both strands by the dideoxy-method (31).

RESULTS AND DISCUSSION

5 independent clones, screened with the human 1st exon, have been analysed. Three fully overlap the human c-myc locus. Two start within the 1st exon and go beyond the 3rd exon. The nucleotide sequence of clone 1, between the SmaI site (position 2226) and the AhaIII site (position 3352) is strictly identical to the one previously published (28). The same result was obtained with the other clones except for clone 4. The sequence derived from that clone shows one single difference, at position 3035 G is replaced by a T. The presence of one difference between clones 1, 2, 3, 5 and clone 4 indicates that we have found the two alleles of the human myc 1st exon of our donor among the 5 isolated λ/myc clones. Unfortunately, the second allele was present only in clone 4 which does not entirely contain the 1st exon but 4/5 of it.

The noticeable fact that emerges from that analysis is the very strong conservation of that sequence. On slightly more than 1000 nucleoti-

des overlapping the human 1st exon there is not a single change for allele a (represented in clones 1, 2, 3, 5) as compared to the previously published sequence, and only one base change in allele b (clone 4) is observed within the 1st intron. This is remarkable because the previous sequence was determined on a DNA of very different origin and a slight polymorphism could have been expected. However, this result fits with the general opinion on the slow rate of evolution of oncogenes.

Six sequences of the 1st human myc exon derived from non pathogenic material have been so far published (11, 13, 15, 22, 25, 28). Only two of them are identical to the sequences presented here (25, 28). To these one can add two sequences derived from the non translocated chromosome of Daudi and Raji cells (19, 21). These last two sequences, which do not span the whole 1st exon, are identical, for the published part, to the sequence of Gazin et al. (28). In the four sequences which differ from the above sequences, the modifications are limited to a few nucleotide changes, but do include deletions or insertions, leading to the lost of its coding capacity (Fig. 1). However one should note that none of these modifications are common to any pair of sequences, which makes it difficult to draw any general rules about the biological significance of this sequence diversity.

From a biological viewpoint, the fact that the 1st myc exon and the sequences surrounding it are highly invariant is probably not meaningless, and indeed supports the various hypotheses concerning the importance of that sequence. The existence of a highly conserved sequence upstream of exons 2 and 3, which code for the classic c-myc protein, fits well for a cis function for that sequence, in which interaction with another nucleotide sequence or with a specific protein regulates expression of the c-myc protein. However, such regulatory interactions are usually restricted to rather short nucleotide sequences and it is difficult to understand why an invariant sequence of one thousand bases would be necessary.

```
          10         20         30         40         50         60         70         80
1  CCCGGGTTCC CAAAGCAGAG GGCGTGGGGG AAAAGAAAAA AGATCCTCTC TCGCTAA-TC TCCGCCCACC GGCCCTTTAT
2  .......... .......... .......... .......... .......... .......C.. .......... ..........
3  .......... .......... .......... .......... .......... .......-.. .......... ..........
4  .......... .......... .......... .......... .......... .......-.. .......... ..........
5  .......... .......... .......... .......... .......... .......-.. .......... ..........

     ***  90         100        110        120        130        140        150        160
1  AATGCGAGGG TCTGGACGGC TGAGGACCCC CGAGCTGTGC TGCTCGCGGC CGCCACCGCC GGGCCCCGGC CGTCCCTGGC
2  .......... .......... .......... .......... .......... .......... .......... ..........
3  .......... .......... .......... .......... .......... .......A-.. .......... ..........
4  .......... .......... .......... .......... .......... .......... .......... ..........
5  .......... .......... .......... .......... .......... .......... .......... ..........

          170        180        190        200        210        220        230        240
1  TCCCCTCCTG CCTCGAGAAG GGCAGGGCTT CTCAGAGGCT TGGCGGGAAA AA-GAACGGA GGGAGGGATC GCGCTGAGTA
2  .......... .......... .......... .......... .......... ..-....... .......... ..........
3  .......... .......... .......... .......... .......... ..-....... .......... ..........
4  .......... .......... .......... .......... .......... ..-....... .......... ..........
5  .......... .......... .......... .......... .......... ..A....... .......... ..........

          250        260        270        280        290        300        310        320
1  TAAAAGCCGG TTTTCGGGGC TTTATCTAAC TCGCTGTAGT AATTCCAGCG AGAGGCAGAG GGAGCGAGCG GGCGGCCGGC
2  .......... .......... .......... .......... .......... .......... .......... ..........
3  .......... .......... .......... .......... .......... .......... .......... ..........
4  .......... .......... .......... .......... .......... .......... .......... ..........
5  .......... .......... .......... .......... .......... .......... .......... ..........

          330        340        350        360        370        380        390        400
1  TAGGGTGGAA GAGCCGGGCG AGCAGAGCTG CGCTGCGGGC GTCCTGGGAA GGGAGATCCG GAGCGAATAG GGGGCTTCGC
2  .......... .......... .......... .......... .......... .......... .......... ..........
3  .......... .......... .......... ..--...... .......... .......... .......... ........CG
4  .......... .......... .......... .......... .......... .......... .......... ........CG
5  .......... .......... .......... .......... .......... .......... .......... ..........

          410        420        430        440        450        460        470        480
1  CTCTGGCCCA GCCCTCCCGC T-GATCCCCC A-GCCAGC-G GTCCGCAACC CTTGCCGCAT CCACGAAA-C TTTGCCCATA
2  .......... .......... .-........ .-......-. .......... .......... ........-. ..........
3  -......... .......... .-........ .-......G. .......... .......... ........A. ..........
4  -......... .......... .-........ .-......G. .......... .......... ........-. ..........
5  .......... .......... .T........ .G.....C-. .......... .......... ........-. ..........

          490        500        510        520        530        540        550        560
1  GCAGCGGGCG GGCACTTT-G CACTGGAACT TACAACACCC GAGCAAGGAC GCGACTCTCC CGACGCGGGG AGGCTATTCT
2  .......... ........AC G......... .......... .......... .......... -......... ..........
3  .......... ........-. .......... .......... .......... .......... -......... ..........
4  .......... ........-. .......... .......... .......... .......... -......... ..........
5  .......... ........-. .......... .......... .......... .........T .......... ..........

          570        580        590        600        610        620        630        640
1  GCCCATTTGG GGACACTTCC CCGCCGCTGC CAGGAC-CCG CTTCTCTGAA AGGCTCTCCT TGCAGCTGCT TAGACGCTGG
2  .......... .......... .......... ......A... .......... .......... .......... ..........
3  .......... .......... .......... ......-... .......... .......... .........C ..........
4  .......... .......... .......... ......-... .......... .......... .........C ..........
5  .......... .......... .......... ......-... .......... .......... .......... ..........

          650    ***  660        670        680        690        700        710        720
1  ATTTTTTTCG GGTAGTGGAA AACCAGGTAA GCACCGAAGT CCACTTGCCT TTTAATTTAT TTTTTTATCA CTTTAATGCT
2  .......... .......... .......... .......... .......... .......... .......... ..........
3  .......... .......... .......... .......... .......... .......... .......... ..........
4  .......-.. .......... .......... .......... .......... .......... .......... ..........
5  .......... .......... .......... .......... .......... .......... .......... ..........
```

Fig. 1. Nucleotide sequence comparison of the sequence overlapping the human c-myc 1st exon from the SMA I site (nucleotide 2226 in 28).
Sequence 1 corresponds to sequence published in (25, 28) and this work, sequence 2 is from (22), sequence 3 from (11), sequence 4 from (15) and sequence 5 is from (13). The ATG at the beginning of the ORF and the TAG stop codon closing this ORF are indicated by stars. Identical nucleotides are indicaded by ., deletion by - .

The existence of an ORF overlapping exon 1 in the five identical sequences reinforces our previous hypothesis that this ORF is used to code for a protein. Also in agreement with this hypothesis are recent

results showing that extracts from various human cells contain a protein specifically recognised by sera raised against synthetic peptides deduced from the sequence of the putative protein (32). The existence of a protein-coding capacity overlapping the 1st myc exon can help explain the necessity for a large highly invariant sequence. However, as for the other hypotheses, it is difficult to explain the very high invariance of the region containing the 1st myc exon on the basis of a protein-coding ORF alone. The ORF could be conserved even if there were several silent point mutations. It is possible that the nucleotide sequence overlapping the human myc 1st exon may be used in several different ways, and that this multiple use imposes constraints which lead to its apparent complete stability.

ACKNOWLEDGMENT

We would like to acknowledge Doctor Alan KAY for his critical reading of the manuscript. This work was supported by CNRS (LP 101) and INSERM (SC 15). SSM was a recipient fellow from ARC.

REFERENCES

1. Land, H., Parada, L.F., and Weinberg, R.A., (1983) Science 222, 771-778.
2. Hann, S.R., Thompson, C.B., and Eisenman, R.N. (1985) Nature 314, 366-369.
3. Kelly, K. , Cochran, B.H., Stiles, C.D., and Leder, P.,(1983) Cell 35, 603-610.
4. Makino, R., Hayashi, K., and Sugimurat (1984) Nature 310, 697-698.
5. Greenberg, M.E., and Ziff, E.B. (1984) Nature 311, 433-437.
6. Watt, R., Stanton, L.W., Marcu, K.B., Gallo, R.C., Croce, C.M., and Rovera, G., (1983) Nature 303, 725-728.
7. Stanton, L.W., Watt, R., and Marcu, K.B. (1983) Nature 303, 401-406.
8. Shih, C.K., Linial, M., Goodenow, M.M., and Hayward, W.S.,(1984) Proc. Natl. Acad. Sci. USA 81, 4697-4701.
9. Persson, H., Leder, P., (1984) Science 225, 718-721.
10. Roussel, M., Saule, S., Lagrou, C., Rommens, C., Beug, H., Graf, T. and Stéhélin, D., (1979) Nature 281, 452-455.
11. Battey, J., Moulding, C., Taub, R., Murphy, W., Stewart, T., Potter, H., Lenoir, G., and Leder, P., (1983) Cell 34, 779-787.
12. Hann, S.R., and Eisenman, R.N. (1984) Mol. Cell. Biol. 4, 2486-2497.
13. Watt, R., Nishikura, K., Sorrentino, J., Ar-Rushdi, A., Croce, C.M., and Rovera, G.,(1983) Proc. Natl. Acad. Sci. USA 80, 6307-6311.
14. Nishikura, K., Ar-Rushdi, A., Erikson, J., Watt, R., Rovera, G., and Croce, C.M. (1983) Proc. Natl. Acad. Sci. USA 80, 4822-4826.
15. Taub, R., Moulding, C., Battey, J., Murphy, W., Vasicek, T., Lenoir, G. and Leder, P., (1984) Cell 36, 339-348.
16. Malanova, Y., Manalov, G., Kieler, J., Leran, A., and Klein G. (1979) Hereditas 90, 5-10.
17. Taub, R., Kirsch, I., Morton, C.C., Lenoir, G.M., Swan, D., Trowick, S., Aaronson, S., and Leder, P., (1982), Proc. Natl. Acad. Sci. USA 79, 7837-7841.

18. Dalla-Favera, R., Martinotti, S., Gallo, R.C., Erikson, S., and Croce, C.M. (1983) Science 219, 963-967.
19. Rabbitts, T.H., Hamlyn, P.H., and Baer, R., (1983) Nature 306, 760-765.
20. Leder, P., Battey, J., Lenoir, G., Moulding, C., Murphy, W., Potter, R.H., Stewart, T., and Taub, R., (1983) Science 222, 765-771.
21. Rabbits, T.H., Forster, A., Hamlyn, P., and Baer, R., (1984) Nature 309, 592-597.
22. Bernard, O., Cory, S., Gerondakis, S., Webb, E., and Adams, J.M.,(1983) EMBO J. 2, 2375-2383.
23. Ar-Rushdi, A., Nishikura, K., Erikson, J., Watt, R., Rovera, G., and Croce, C.M.,(1983) Science 222, 390-393.
24. Dunnick, W., Shell, B.E., and Dery, C., (1983) Proc. Nalt. Acad. Sci. USA 80, 7269-7273.
25. Saito, H., Hayday, A.C., Wiman, K., Hayward, W.S., and Tonegawa, S., (1983) Proc. Natl. Acad. Sci. USA 80, 7476-7480.
26. Rabbits, P.H., Forster, A., Stinson, M.A., and Rabbits, T.H., (1985) EMBO J. 4, 3727-3733.
27. Eick, D., Piechaczyk, M., Henglein, B., Blanchard, J.M., Traub, B., Kofler, E., Wiest, S., Lenoir, G.M., and Bornkamm, G.W., (1985) EMBO J. 4, 3717-3725.
28. Gazin, C., Dupont de Dinechin, S., Hampe, A., Masson, J.M., Martin, P., Stéhélin, D., and Galibert, F., (1984) EMBO J. 3, 383-387.
29. Maniatis, T., Fristsch, E.F., and Sambrook, J., (1982) In Molecular Cloning. A laboratory manuel.
30. Efimov, V.A., Buryakova, A.A., Reverdatto, S.V., Chakhmakhcheva, O.G., and Ovchinnikov, Y.A., (1983) Nucleic Acids Res. 11, 8369-8387.
31. Sanger, F., Nicklen, S., and Coulson, A.R., (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467.
32. Gazin, C., Rigolet, M., Briand, J.P., Van Regenmortel, M.H.V., and Galibert, F., EMBO J. (in press).