Nucleic Acids Research, Vol. 20, No. 5 ***1087–1091***

Evolutionary variation of the CCAAT-binding transcription factor NF-Y

Xiao-Yan Li, Roberto Mantovani, Rob Hooft van Huijsduijnen, Isabelle Andre, Christophe Benoist* and Diane Mathis
Laboratoire de Génétique Moléculaire des Eucaryotes du CNRS, Unité 184 de Biologie Moléculaire et de Génie Génétique de l'INSERM, Institut de Chimie Biologique, Faculté de Médecine, Strasbourg, France

Received November 21, 1991; Revised and Accepted February 3, 1992 EMBL accession nos+

ABSTRACT

NF-Y is a CCAAT-specific transcription factor thought to be involved in the regulation of a variety of eukaryotic genes. It shows a striking sequence similarity with the yeast factor HAP2/3. In an attempt to trace back its evolutionary history, we succeeded in isolating NF-Y cDNA clones from a plant and from several species of vertebrates. The patterns of sequence conservation delineate potential functional domains: A central, highly conserved, domain likely responsible for DNA-binding and subunit interaction; more evolutionarily flexible flanking regions, in which variability is clustered, individualizing conserved glutamine or acidic amino-acids putatively involved in protein-protein contacts.

INTRODUCTION

The CCAAT box is a common cis-acting element found in the promoter and enhancer regions of a large number of genes in higher eukaryotes. Its role as a positive promoter element was defined via mutational analysis of several genes in many eukaryotic species (for review, see refs 1,2).

Diverse DNA binding proteins have been reported to bind to CCAAT boxes, either with exquisite specificity or with a loose specificity encompassing the CCAAT motif (for refs, see refs 1,2). Among these is the factor variously called CBF, NF-Y or CP1 (hereafter referred to as NF-Y), which seems to have an absolute requirement for the CCAAT pentanucleotide (3–10). It is composed of two subunits (NF-YA and NF-YB), and is present in all murine tissues so far examined.

Binding activities similar to NF-Y have been described in a variety of eukaryote orders: primates, rodents, birds, echinoderms (3–15). In addition, the yeast Saccharomyces cerevisiae harbors two genes (HAP2 and HAP3) that code for a heterodimeric transcription factor which recognises regulatory elements of cytochrome genes induced by non-fermentable carbon sources (16,17). The DNA motif recognized by the HAP activator is virtually identical to the CCAAT box. HAP2 and HAP3 are also able to heterodimerize with the subunits of NF-Y (9). It did not, then, come as much of a surprise that the recent cloning of the cDNAs coding for the two subunits of NF-Y demonstrated a striking sequence homology between NF-YA/B and HAP2/3 (18–21). For murine NF-YB, for example, a large central domain was perfectly colinear with and 73% identical to the central region of HAP3 (18,20). The flanking segments, on the other hand, showed no homology between the mouse and yeast genes.

In light of this striking evolutionary conservation, and to gain further information on the structure/function relationship of NF-YB, we decided to trace back the phylogenetic history of this transcription factor, choosing a number of species placed at different levels of the evolutionary tree of eukaryotes (Figure 1). We report here the sequencing of the NF-YB equivalents in man, chicken, toad, lamprey and maize.

MATERIALS AND METHODS

The PCR-based strategy we used here to isolate maize NF-YB was originally described in ref 18. PCR primers were chosen to encompass the stretches of greatest homology in the HAP3/NF-YB homology region. Primers were end-labelled with ^{32}P. The substrate for the amplification was whole DNA from a lambda gt11 cDNA library derived from RNA of young maize plants (Zea mais) (a kind gift from C. Gigot's lab). PCR products were run on denaturing acrylamide gels, and bands of the expected size were eluted and chemically sequenced (Maxam and Gilbert, 1980). PCR primers used for these amplifications were: 5'TC-C/TTCNCCG/ATTG/A/TATNGTC/TT3' and 5'AAA/GGA-A/GTGT/CGTNCAA/GGAA/GTG-3'. To isolate NF-YB cDNA clones from other species, we used standard screening methods with low stringency hybridization, using a probe derived from the most conserved domain of mouse NF-YB, from the

* To whom correspondence should be addressed

+ X55315, X55316, X59703, X59710–X59714 (incl.)

clone YB-EM38 (18). The chicken (Gallus gallus) and toad (Xenopus laevis) libraries were carried in the λZAP vector, while the lamprey (Petromyzon marinus) library was in λgt11 (kind gifts from M. Cooper, L. du Pasquier, and G. Littman, respectively). In spite of repeated screening attempts, including the use of cognate probes, we were unable to isolate longer clones from the chicken or toad libraries. cDNA inserts were converted or subcloned into phagemid vectors, and single-stranded templates derived therefrom sequenced by the dideoxynucleotide technique, as described (18). All sequences reported here were obtained from both DNA strands and, except for short stretches, derived from two or more independent clones.

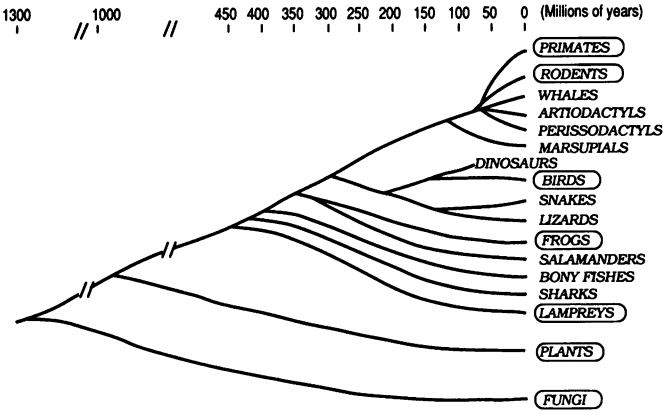

Figure 1. This simplified evolutionary tree positions the species (circled) analyzed in this study. Schematic adapted from refs 27, 28.

RESULTS

We isolated NF-YB cDNA clones from several species by a combination of PCR-based cloning and of direct screenings of cDNA libraries at reduced stringency. For the PCR cloning, we used the strategy previously employed for the original cloning of NF-YB (direct cloning of large fragments obtained after 'anchored PCR' did not prove successful) (18). Short amino-acid clusters of absolute identity between yeast and mouse were located and degenerate PCR primers corresponding to these stretches synthesized. One of these primers was 5'-end labeled with ^{32}P, and PCR amplification was performed on a template consisting of an entire cDNA library from a given species. An amplified fragment of the expected size was isolated, and sequenced using the Maxam and Gilbert method. The unambiguous sequence between the primers was then used for a standard screening of the cDNA library. As it turned out, this strategy was essential only for maize. For all other species examined, the conservation of NF-YB was high enough to allow direct screening of lambda-carried libraries at moderate stringency. NF-Y cDNA clones were rare (roughly 1 in 10^5) in all species.

Inserts from the cDNA clones were subcloned into pBluescript for sequencing. In most instances, the sequences we report derive from several independent cDNA clones (human, 2; chicken, 5; toad, 1; lamprey, 2; maize, 2). Sequences were determined on both strands. The protein protein sequences derived from these data are shown in Figure 2 and are discussed below. The nucleotide sequence data, not shown, is available from the authors or from databanks.

We did not succeed in identifying PCR fragments or in isolating clones from D. melanogaster cDNA libraries. While this failure

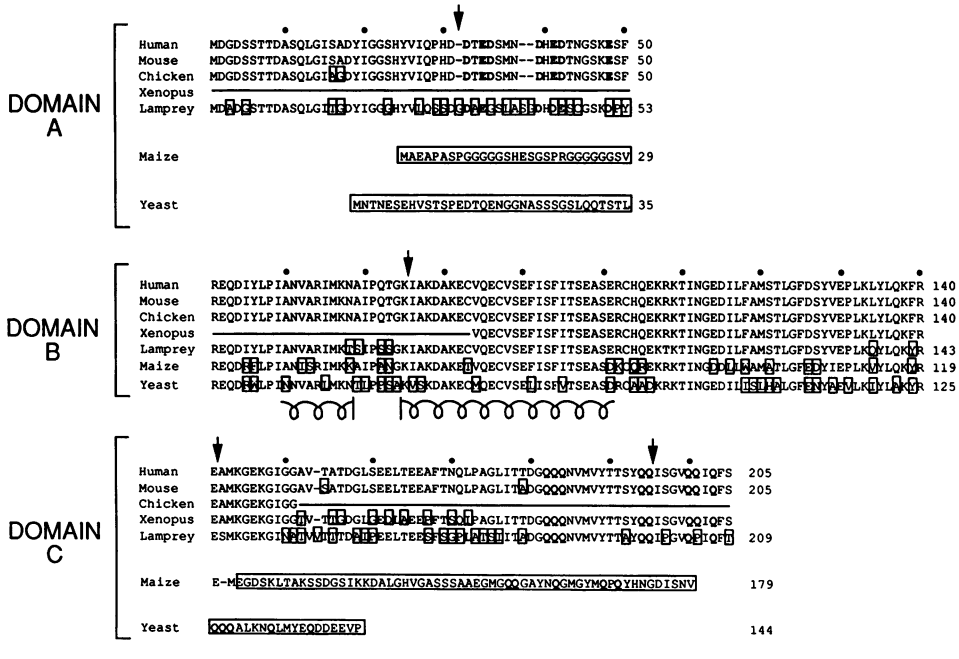

Figure 2. NF-YB protein sequences. The sequences are aligned with the human NF-YB, and are split into three domains: the central 'B' domain corresponds to the domain of colinearity and strong sequence homology between murine NF-YB and yeast HAP3; 'A' and 'C' domains are the N and C-terminal flanking regions, which show no such similarity between yeast and mouse (no attempt is made to align maize and yeast sequences with the others in the A and C domains). Missing stretches of chicken or Xenopus sequences are shown as a solid line. Amino-acid differences relative to the human protein are boxed. Bold characters denote the acidic or glutamine residue in the A and C domains, respectively. Vertical arrows point to the positions of introns, as determined elsewhere (22). The EMBL database accession numbers for these sequences are in the order of the figure, X59710, X55316, X59713, X59703, X59712 and X59714, for yeast HAP sequences, see ref 17 and refs therein.

certainly does not constitute proof that an NF-YB homologue is absent, it is in keeping with our observation that nuclear extracts from Drosophila cell lines and embryos do not contain NF-Y-like CCAAT-binding activity; in contrast, such an activity is readily detectable in extracts from plant cells; for example, a proportion of batches of commercially available wheat-germ cell-free extracts contain an NF-Y like DNA-binding activity; (R.H., unpublished). If true, this observation would suggest that gene regulation systems in Drosophila have forsaken a control motif otherwise used universally in eukaryotes.

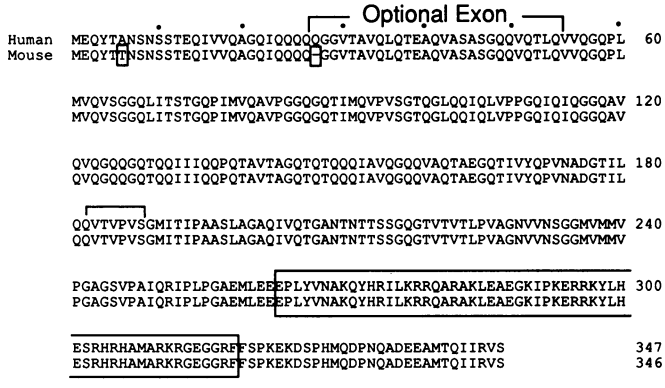

Figure 3. NF-YA protein sequences. The two differences between human and murine sequences are boxed on the mouse sequence. A bracket delineates the optional 28 a.a. stretch of exon B. The domain of high sequence similarity between yeast and mammalian sequences is boxed. The EML database accession numbers for these sequences are: X59711 and X55315.

Although we did not intend to perform an extensive analysis in this instance, we also determined the complete sequence of two human NF-YA cDNA clones, shown in Figure 3. Where they overlap, our sequence is in perfect agreement with the partial sequence determined by Becker absence of an 84 bp stretch which corresponds exactly to the alternative splicing of NF-YA mRNA that we have described for the mouse gene. In mouse, this variation corresponds to an optional splicing of exon B, and shows marked tissue-specific bias (22).

DISCUSSION

Evolution of NF-Y

A number of general points can be made from the examination of the sequences of Figures 1 and 2:

1) NF-Y is, as a whole, a very strongly conserved protein: human and murine NF-YA differ by only two amino-acids, as do human and murine NF-YB. This quasi-invariance also extends to chicken, since human and chicken NF-YB show only two (very conservative) changes out of 150 amino-acids. It is only when genetic distances exceed 250 MY that one begins to see significant divergence, predominantly clustered in the A and C domains.

This high degree of conservation reflects selection pressure acting at the protein level, since amino-acid sequences are better conserved than nucleotide sequences (99% vs 89% between human and murine NF-YA, for example), reflecting a large number of silent nucleotide substitutions. This high degree of conservation may stem from the fact that NF-Y interacts with a number of DNA targets as well as other transcription factors. Further, since NF-Y appears to be encoded by unique genes, not members of multigene families, it should be rather resistant

Figure 4. Sequence substitutions between human and murine transcription factors. Each transcription factor is schematized by a horizontal box, drawn to scale, in which the DNA-binding domain is stippled. Individual amino-acid substitutions or single deletions are shown as vertical black bars. Insertions or deletions are shown as a single bar, the number inside the inverted triangle indicating the size of the insertion/deletion. Sequence data for the comparison originated from the following references: Histone H4, refs 29, 30; Oct 2^a , 31 and refs therein; TFIID, 32 and refs therein; ets, 33, 34 and refs therein; TCF1, 35, 36; jun, 37 and refs therein, fos, 38,39; NF-?B, 40,41; rel, 42,43; myc, 44,45.

to the 'molecular drive' evolutionary forces operating on multigene families (eg. gene conversion, unequal crossing-over; for review, see ref, 23).

2) Conservation between kingdoms is confined to the central B domain, with very sharp boundaries that are identical in all pairwise comparisons between animal, plant or yeast sequences. Maize NF-YB is closer to its animal counterpart (77% identity) than it is to the yeast equivalent (62% identity). We and others have hypothesized that this strong conservation probably corresponds to the subunit-interaction and DNA-binding functions (18−21). This prediction has been supported experimentally for yeast (17) and mouse (U. Pessara, unpublished). Since NF-Y binds to CCAAT boxes in a number of genes in animals (and presumably several in yeast), it is clear that tolerance to evolutionary drift in the DNA-binding domain must be very low.

3) The A and C domains show no conservation between kingdoms, except perhaps for a high concentration of glutamine residues at the end of the C domain. On the other hand, they are very highly conserved within vertebrates. We interpret this conservation to reflect the several interactions that NF-Y must engage in: NF-Y cooperates with different additional factors for the regulation of albumin, MHC or actin genes (amongst others). This multiplicity of interactions must place serious constraints on the evolutionary potential of the protein/protein contact surfaces. Coevolution by compensating mutations, theoretically possible with protein pairs that only interact with each other, would with each other, would seem difficult in this instance.

We have recently shown that the Gln-rich region of NF-YA and the acidic region of NF-YB act as transcriptional activators (22 and R.M., unpublished). The sequence conservation we observe in these regions is at odds with the notion that acidic or Gln-rich activators have few sequence constraints beyond a high proportion of Gln or acidic amino-acids (1,24,25). To the contrary, our data point to a specific structural organization, highly constrained and sensitive to even very conservative changes. The root of the discrepancy probably lies in the artificial nature of the chimeric constructs used in transcription activation assays.

NF-Y Structure/Function

We had hoped that an evolutionary analysis of NF-Y might give us clues about functional domains. While the human and chicken sequences are obviously uninformative in this respect, several points can be made from the comparative analysis of other NF-YB sequences:

The A domain (a.a. 1−50). The comparison between the most distant animal organisms (man and lamprey) reveals a stronger conservation in the first 26 amino acids. We can thus tentatively divide this domain into two: A1 (1−26) and A2 (26−50). A2 is clearly more divergent than A1, even showing a length difference between human and lamprey NF-YB. A2 also corresponds to the acidic stretch; interestingly, the acidic positions are preferentially preserved. Note that an intron maps close to the border between A1 and A2 (Figure 2).

The B domain (a.a. 51−140). The DNA-binding domain is virtually identical in all animal species. Most of the variation between lamprey and human NF-YB is clustered, and coincides with a short stretch between two marked α-helices predicted by secondary structure analysis (positions 68−73); an intron also coincides with this junction (see figure 2 and ref 22).

We find a mutation in maize NF-YB that replaces by a Thr the otherwise conserved Cys-85. This finding would appear to render unlikely our previously proposed 'half-finger' model, according to which Cys-85 and Cys-89 form a tetrahedral metal coordination complex together with two His residues from NF-YA (18). Yet it is interesting to note that Cys→Thr is a tolerated change in the similar protein-metal-protein complex that links CD4 to p56lck (D. Littman, personal communication).

The C domain (a.a. 141−205). Again the toad and lamprey sequences are the most informative, and their comparison with human and murine NF-YB clearly splits the C domain into two. The first 39 amino acids (142−180; C1) show a considerable (for NF-YB) divergence between man and toad or lamprey, with no particular striking sequence feature. C2 (180−205), on the other hand, diverges very little. It is markedly Gln-rich, with an absence of charged amino-acids that is characteristic of Gln-rich transcriptional activation regions in other factors (1 and refs therein). It seems likely that this conserved C2 stretch is involved in some form of protein-protein interaction, possibly even NF-Y dimerization as described for SP1 (26; that NF-Y multimerizes is suggested by the apparent size of the protein-DNA complex, relative to the size of NF-YA and NF-YB—see refs 6, 18—and by preliminary deletion experiments (U. Pessara unpublished data).

NF-Y and other transcription factors

We thought it interesting to compare the high degree of conservation we observe for NF-Y with that found for other transcription factors. As shown on Figure 4, transcription factors tend to be rather conserved overall: 0.5 to 8.8% amino-acid substitutions between human and murine homologs. For comparison, human and murine homologs for other types of proteins show the following range of amino-acid substitution frequencies: interleukin-2, 35.5%; major histocompatibility complex class II (Ea vs DRa), 27.3%; lysozyme, 21.6%; terminal-deoxynucleotidyl-transferase, 18.3%; cytochrome c, 7.5%. The high conservation of transcription factors probably reflects their role in a key cellular process, as well as the combinatorial strategy of eukaryotic gene regulation. Just like NF-Y, the DNA binding domain in most transcription factors is by far the best preserved evolutionarily. With hardly an exception, substitutions in the human/mouse comparisons of Figure 4 fall outside the DNA binding domains.

Among transcription factors, NF-YA and NF-YB are the most conserved between human and murine sequences, with the same rate of substitution as the quasi-invariant histone H4, and lower than that of the 'general transcription factor' TFIID. This observation further substantiates the notion that NF-Y plays a broad role in regulating a large number of eukaryotic genes.

ACKNOWLEDGEMENTS

We are grateful to Drs. J. Klein and J. Kaufman for critical reading of the manuscript, to P. Gerber for assistance, and to A. Staub and F. Ruffenach for oligonucleotide synthesis. This work was supported by institutional grants from the CNRS and the INSERM, and by the Association pour la Recherche contre le Cancer. RM, XYL and RH received fellowships from the EMBO, the Université Louis Pasteur, and the Fondation pour la Recherche Médicale, respectively.

REFERENCES

1. Mitchell, P.J., and Tjian. R. (1989). Science 245,371−378.
2. Johnson, P.F., and McKnight. S.L. (1989). Annu. Rev. Biochem. 58,799−839.
3. Oikarinen, J., Hatamochi, A. and de Crombrugghe. B. (1987). J. Biol. Chem. 262,11064−11070.
4. Hatamochi, A., Golumbek, P.T., Van Schaftingen, E. V. and de Crombrugghe. B. (1988). J. Biol. Chem. 263,5940−5947.
5. Dorn, A., Bollekens, J., Staub, A., Benoist, C. and Mathis. D. (1987). Cell 50,863−872.
6. Hooft van Huijsduijnen, R., Bollekens, J., Dorn, A., Benoist, C. and Mathis. D. (1987). Nucleic Acids Res. 15,7265−7272.
7. Raymondjean, M., Cereghini, S. and Yaniv. M. (1988). Proc. Natl. Acad. Sci. USA. 85,757−761.
8. Wuarin, J., Mueller C. and Schibler. U. (1990). J. Mol. Biol. 214,865−874.
9. Chodosh, L.A., Olesen, J. Hahn, S., Baldwin, A.S., Guarente, L. and Sharp. P.A. (1988). Cell 53,25−35.
10. Chodosh, L.A., Baldwin, A.S., Carthew, R.W. and Sharp. P.A. (1988). Cell 53,11−24.
11. Barberis, A., Superti-Furga, G. and Busslinger. M. (1987). Cell 50,347−359.
12. Knight, G.B., Gudas, J.M. and Pardee. A.B. (1987). Proc. Natl. Acad. Sci. USA 84,8350−8354.
13. Quitschke W.W., Lin, Z-Y., DePonti-Zilli, L. and Paterson. B.M. (1989). J. Biol. Chem. 264,9539−9546.
14. van Wijnen, A.J., Massung, R.F., Stein, J.L. and Stein. G.S. (1988). Biochemistry 27,6534−6541.
15. Kim, C.G., and Sheffery. M. (1990). J. Biol. Chem. 265,13362−13369.
16. Hahn, S. and Guarente. L. (1988). Science 240,317−321.
17. Olesen, J.T., and Guarente. L. (1990). Genes & Development. 4,1714−1729.
18. Hooft van Huijsduijnen, R., Li, X-Y., Black, D., Matthes, H., Benoist, C. and Mathis. D. (1990). EMBO J. 9,3119−3127.
19. Maity, S.N., Vuorio, T. and de Crombrugghe. B. (1990). Proc. Natl. Acad. Sci. USA 87,5378−5382.
20. Vuorio, T., Maity, S.N. and de Crombrugghe. B. (1990). J. Biol. Chem. 265,22480−22486.
21. Becker, D.M., Fikes, J.D., Guarente. L. (1991). Proc. Natl. Acad. Sci. 88,1968−1972.
22. Li, X-Y., Hooft van Huijsduijnen, R., Mantovani, R., Benoist, C, Mathis, D. J; Biol. Chem, in press.
23. Dover, G.A. (1986). TIG 2,159−165.
24. Sigler, P.B. (1988). Nature 333,210−212.
25. Ptashne, M. (1988). Nature 335,683−689.
26. Su, W., Jackson, S., Tjian, R. and Echols. H. (1991). Genes & Development 5,820−826.
27. Nei, M. (1987). Molecular Evolutionary genetics, Columbia University Press, NY.
28. McLaughlin, P.J. and Dayhoff. M.O. (1970). Science 168,1469−1470.
29. Sierra, F., Stein, G., Stein. J. (1983). Nucl. Acids. Res. 11,7069−7086.
30. Seiler-Tuyns, A., Birnstiel. M.L. (1981). J. Mol. Biol. 151,607−625.
31. Hatzopoulos, A.K., Stvoykova, A.S., Erselius, J.R., Goulding, M., Neuman, T. and Gruss. P. (1990). Development 109,349−362.
32. Tamura, T-A., Sumita, K., Fujino, I., Aoyama, A., Horikoshi, M., Hoffman, A., Roeder, R.G., Muramatsu, M. and Mikoshiba. K. (1991). Nucl. Acids Res. 19,3861−3865.
33. Watson, D.K., McWilliams, M.J., Lapis, P., Lautenberger, J.A., Schweinfest, C.W. and Papas. T.S. (1988). Proc. Natl. Acad. Sci. 85,7862−7866.
34. Gunther, C.V., Nye, J.A., Bryner, R.S. and Graves. B.J. (1990). Genes and Develop. 4,667−679.
35. Travis, A., Amsterdam, A., Belanger, C. and Grosschedl. R. (1991). Genes and Development. 5,880−894.
36. van de Wetering, M., Oosterwegel, M., Dooijes D. and Clevers. H. (1991). EMBO J. 10,123−132.
37. Ryseck, R-P., Hirai, S.I., Yaniv, M. and Bravo. R. (1988). Nature 334,535−537.
38. Van Straaten, F., Muller, R., Curran, T., Van Beveren, C., Verma. I.M. (1983). Proc. Natl. Sci. U.S.A. 80,3183−3187.
39. Van Beveren, C., Van Straaten, F., Curran, T., Muller, R., Verma. I.M. Cell 32,1241−1255.
40. Ghosh, S., Gifford, A.M., Riviere, L.R., Tempst, P., Nolan. G.P., Baltimor, E.D. (1990). Cell 62,1019−1029.
41. Kieran, M., Blank, V., Logeat, F., Vandekerckhove, J., Lottspeich, F., LeBail, O., Urban, M.B., Kourilsky, P., Baeuere, P.A. and Israel, A. (1990). Cell 62,1007−1018.
42. Grumont, R.J. and Gerondakis. S. (1989). Oncogene Res. 4,1−8.
43. Brownell, E., Mittereder, N. and Rice. N.R. (1989). Oncogene 4,935−942.
44. Bernard, O., Cory, S., Gerondakis, S., Webb, E. and Adams. J.M. (1983). EMBO J. 2,2375−2383.
45. Battey, J., Moulding, C., Taub, R., Murphy, W., Stewart, T., Potter, H., Lenoir, G. and Leder. P. (1983). Cell 34,779−787.