Volume 335, number 2, 289–293 FEBS 13341 December 1993
© 1993 Federation of European Biochemical Societies 00145793/93/$6.00

Evolutionary clustering and functional similarity of RNA-binding proteins

Kaoru Fukami-Kobayashi, Shirou Tomoda, Mitiko Gō*

Department of Biology, Faculty of Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-01, Japan

Received 11 September 1993

RNA-binding proteins (RNPs) involved in splicing, processing and translation regulation contain one to four RNA-binding domains. We constructed a phylogenetic tree for the RNA-binding domains, including those of poly(A)-binding protein (PABP), splicing factors, chloroplast RNPs, hnRNPs, snRNP U1-70K, nucleolin and *Drosophila* sex determinants. Proteins with similar functions were found to have closely related RNA-binding domains and common domain organizations. In light of these observation, one can assume the function of an RNA-binding protein, based on the evolutionary relationship between its RNA-binding domain(s) and domain organization, as compared with other RNPs.

RNA-binding domain; Splicing factor; Ribonucleoprotein; Poly(A)-binding protein; *Drosophila* sex determinant; Gene duplication

1. INTRODUCTION

Biological functions of proteins that have an RNA-binding domain(s) are of a wide range; these domains are involved in splicing, processing, translation of RNA, and other protein–RNA interactions (for review see [1,2]). The RNA-binding domain consists of about 80 amino acid residues. Evidence obtained suggested that the minimal region essential for RNA-binding is mainly occupied by an RNA-binding domain [3–5]. X-Ray crystallographic and/or NMR solution analyses of the RNA-binding domains of snRNP U1A [6,7] and hnRNP C [8] show that the domain forms a globular structure. The RNA-binding domain is, therefore, a functional and structural unit of the RNA-binding proteins.

The RNA-binding domain is found in various eukaryotic nuclear and cytosolic RNA-binding proteins, thus, the RNA-binding domain is a common functional and structural unit shared among proteins that interact with RNA. The RNA-binding domains include conserved motifs RNP-1 (or RNP-CS) and RNP-2. The three-dimensional structures of snRNP U1A and hnRNP C are similar to each other in spite of the low identity in their overall amino acid sequences (< 20%). Most of the conserved residues are hydrophobic ones which constitute hydrophobic cores shared in the RNA-binding domains. These facts suggest that the RNA-binding domains were derived from a common ancestral gene. The RNA-binding domain is also found in chloroplast proteins [9–11]. Although these chloroplast proteins are encoded in the nuclear genome, many are also considered to used to be encoded in the chloroplast genome [12]. On the basis of the symbiotic origin of chloroplasts, the RNA-binding domain might have occurred prior to the divergence of eukaryotes and prokaryotes. It seems important, therefore, to analyze the evolutionary relationship of the RNA-binding domains of those proteins in order to better understand the origin and molecular evolution of various functions of the RNA-binding proteins.

While the amino acid sequences of the RNA-binding domains include tightly conserved stretches, such as RNP-1 (or RNP-CS) and RNP-2, they are loosely conserved, as whole sequences. In addition, sequences of the RNA-binding domains are not long enough to compute accurately the evolutionary distances among them. Thus, it has been difficult to clarify the evolutionary relationship of RNA-binding domains. One of us (K.F.-K.) has recently developed a new method for estimating the evolutionary distance between the short amino acid sequences with low identity [13]. We used this method to construct a phylogenetic tree of RNA-binding domains.

We analyzed sequences of some RNA-binding proteins to address the following: (1) RNA-binding domains are often tandemly repeated. How early were those repeated structures established? (2) Does the evolutionary relationship among the RNA-binding domains reflect functional similarity of the proteins bearing these domains? (3) The RNA-binding proteins also have auxiliary domains unique to each type of protein [14]. Did the auxiliary domain(s) of an RNA-binding protein diverge independently of or together with the RNA-binding domain(s) of the protein?

*Corresponding author. Fax: (81) (52) 782-9609.

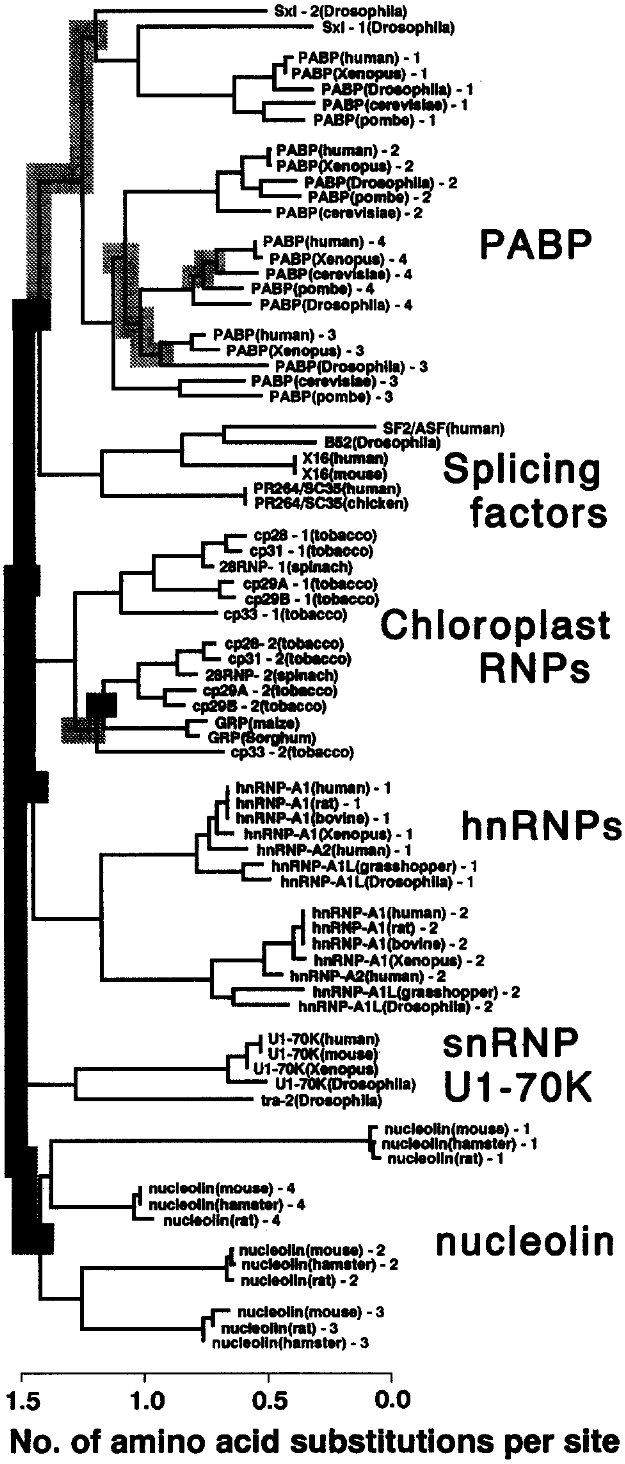
2. MATERIALS AND METHODS

We used the amino acid sequences of 73 RNA-binding domains to construct a phylogenetic tree, each forms part of 35 amino acid sequences of RNA-binding proteins or was deduced from nucleotide sequences. The amino acid sequence of each domain has sufficient identity with other sequences (> 15%) to estimate reliable evolutionary distances, using the similarity distance (SD) method [13]. The RNA-binding domains of snRNP U1A and hnRNP C, whose three-dimensional structures have been resolved, could not be included into the estimation of evolutionary distance because their identity with some of other sequences is lower than 15%. Data on the accepted point mutations and frequencies for the 20 amino acid residues [15] were used for the probability model of the SD method. A phylogenetic tree was constructed by the neighbor-joining (NJ) method [16]. We also carried out bootstrap resampling [17] to test the reliability of each branch in the NJ tree.

3. RESULTS

Fig. 1 shows a tree of the 73 RNA-binding domains from 35 sequences. Of these, 22 have the repeated structure of the RNA-binding domains. The RNA-binding domains of poly(A)-binding protein (PABP), chloroplast RNPs, hnRNPs, including A1, A2 and their homologous proteins, and nucleolin clustered respectively in the constructed tree. This means that repeated structures of the RNA-binding domains, i.e. the four repeats in PABP and nucleolin and the two repeats in chloroplast RNPs and hnRNPs were established by duplication, which occurred independently of these proteins.

In PABPs, each of the four RNA-binding domains derived from the five species, from yeasts and animals, clustered in the phylogenetic tree. Successive duplications that led to the repeated structure of the RNA-binding domains in PABP, therefore, preceded the divergence of yeasts and animals. Nucleolin also has the four-repeated structure of the RNA-binding domains. The structure appeared independently of the one in PABP. The hnRNP A1, A2 and A1-like proteins have two RNA-binding domains. The tree indicates that the ancestral gene of the hnRNPs had the two RNA-binding domains long before the divergence of vertebrates and invertebrates, and has diverged to the present hnRNP genes while preserving the repeated structure. In the ancestral gene of the six chloroplast RNPs, the RNA-binding domain was duplicated and formed the tandem repeat, independently of the hnRNPs; the re-



→

Fig. 1. Phylogenetic tree of RNA-binding domains. The branches whose occurrence was less than 50% and 25% in the bootstrap 500 trials are shown with light and dark shade, respectively. The RNA-binding proteins whose domain(s) was used in the tree construction are sex-lethal (Sxl) [41], poly(A) binding protein (PABP) [27,28,42–45], alternative splicing factor (SF2/ASF) [29,30], B52 [31,32], X16 [18,33], PR264/SC35 [34,35], chloroplast RNPs from tobacco (cp28, cp29A, cp29B, cp31, cp33) [9,10] and from spinach (28RNP) [11], glycine-rich protein (GRP) [46,47], hnRNP A1 [36,37,40,48], hnRNP A1-like proteins (hnRNP-A1L) [38,39], hnRNP A2 [19], U1 snRNP 70K protein (U1-70K) [49–53], transformer-2 (tra-2) [54] and nucleolin [55–57]. When more than one RNA-binding domain are contained in a protein, their numbers are assigned starting at the N-terminal domain. Glycine-rich proteins (GRP) are probably cytosolic proteins, although their RNA-binding domains belong to a cluster designated 'chloroplast RNPs' in the tree.

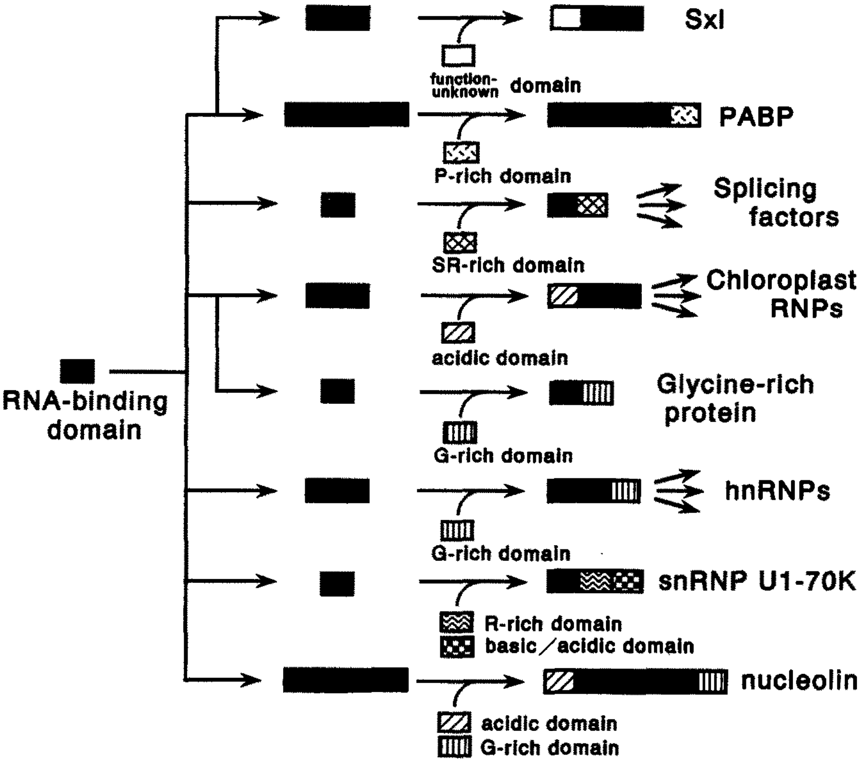


Fig. 2. Evolutionary scheme of the RNA-binding proteins.

peat has been conserved during evolution of the chloroplast RNPs, as was also suggested by Ye et al. [10].

The tree also shows that proteins with similar functions have closely related RNA-binding domains. The RNA-binding domains of SF2/ASF, B52, X16 and PR264/SC35, all of which are splicing factors [18], make a cluster. The RNA-binding domains of hnRNP A1 and A2 and the *Drosophila* A1-like protein, which are components of the hnRNP particle [19–21], also form a cluster in the tree.

Correspondence of the molecular function of the RNPs to the evolutionary relationship of their RNA-binding domains is remarkable in Sxl and tra-2, two of the sex determinants of *Drosophila.* Both belong to a hierarchy of regulatory genes for somatic sex determination and control alternative splicing of mRNA of their own and their downstream protein; however, mechanisms controlling splicing differ. The product of *tra-2* induces the female-specific splicing together with product of *tra* by recruiting general splicing factors to a regulatory element located downstream of a female-specific 3′ splice site [22]. The RNA-binding domain of tra-2 forms a cluster with those of snRNP U1-70K, a component of the spliceosome. The product of *Sxl*, in contrast, inhibits the non-sex-specific splicing by recognizing the uridine octamer at the acceptor site of non-sex-specific splicing [23]. The RNA-binding domains of Sxl are most closely related to those of PABP. Both of Sxl and PABP bind to specific nucleotide sequences.

4. DISCUSSION

The phylogenetic tree indicates that the repeated structures of the RNA-binding domains appeared in the early stage of biological evolution, and have been conserved for a long time. In PABP, establishment of the repeated structure preceded the divergence of animals and yeasts; in hnRNP, that is far before the divergence of vertebrates and invertebrates; and in chloroplast RNPs, that might have been prior to the integration of chloroplast to the ancestral cell of eukaryotes. This is quite different from evolution of the repeated structure of the Ig- or EGF-like domains in animal extracellular proteins and blood coagulation factors [24–26], where a newly created repeated structure is often observed, in some species. The highly conserved repeated organization of the RNA-binding domain implies that the each domain has its own unique functional role. Indeed, each of the four domains of *Xenopus* and yeast PABP has its own RNA-binding activity and specificity [27,28].

Evolutionary clustering of the RNA-binding domains suggests that function of the tobacco chloroplast RNPs (cp28, cp29A, cp29B, cp31 and cp33), which is unknown, is similar to that of the chloroplast 28 kDa RNA-binding protein from spinach (28RNP). The protein from spinach is required for the processing and/or stability of plastic mRNA 3′-end [11]. The tobacco RNPs may be involved in maturation of chloroplast pre-mRNA. Similarly, grasshopper hnRNP A1-like protein is probably a component of the hnRNP particle, because the tree shows that its organization of the repeated RNA-binding domains has evolved from a common ancestor with authentic hnRNPs.

The RNA-binding proteins have additional domains which are characterized by an abundance of specific residues in their amino acid sequences [14]. The splicing factors SF2/ASF, B52, X16 and PR264/SC35, whose RNA-binding domains make a cluster in the tree, have serine-arginine-rich (SR-rich) domains in their C-terminal regions [18,29–35] (Fig. 2). This implies that the fusion of the RNA-binding domain and the SR-rich domain occurred before the divergence of splicing factors, and that the SR-rich domain has evolved together with the RNA-binding domain thereby conserving the fused domain organization. HnRNPs share the glycine-rich (G-rich) domain on the C-terminal side of their two-repeated structure of the RNA-binding domains [19,36–40]. This also suggests evolutionary conservation of the domain organization of their common ancestor. An evolutionary pathway of proteins that share the RNA-binding domains is proposed (Fig. 2), taking into account duplication and divergence of the RNA-binding domains, as well as structural arrangements of auxiliary domains. The ancestral RNA-binding domain may have appeared even before the divergence of eukaryotes and prokaryotes, because the RNA-binding domains are also found in chloroplast proteins. Repeated duplication of the RNA-binding domain and addition of auxiliary domains occurred, independently in PABP, splicing factor, chloroplast RNP, hnRNP, snRNP U1-70K and nucleolin, before the appearance of primordial eukaryotes. The glycine-rich (G-rich) domains of hnRNP, nucleolin and glycine-rich protein (GRP) do not seem to be homologous because the sequence patterns of the domains are different from each other except for an abundance of glycine residues. The functional role of each domain was assigned independently to these proteins and their domain organizations have been conserved. The auxiliary domain(s) shared by proteins with a similar function diverged together with the RNA-binding domain(s). The two exceptions are *Drosophila* Sx1 and plant glycine-rich protein (GRP). Although establishment of the repeated structure of the RNA-binding domains in Sx1 occurred together with that in PABP, the number of repeated RNA-binding domains and fusion of an auxiliary domain with the repeated domains of Sx1 differ from those of PABP. The RNA-binding domain of the glycine-rich protein derived from the C-terminal RNA-binding domain of chloroplast RNP (Fig. 1). Fusion of the RNA-binding domain and a G-rich domain produced plant glycine-rich protein, while fusion of an acidic domain and two-repeated RNA-binding domains produced chloroplast RNP.

Acknowledgements: We thank Professor M. Sugiura for introduction to RNPs and for various discussion, Drs. N. Saitou and T. Imanishi for programs for computing and drawing an evolutionary tree, and M. Ohara for critical comments on the manuscript. This research was supported in part by the Ogasawara Foundation for the Promotion of Science & Engineering.

REFERENCES

[1] Mattaj, I.W. (1993) Cell 73, 837–840.

[2] Kenan, D.J., Query, C.C. and Keene, J.D. (1991) Trends Biochem. Sci. 16, 214–220.

[3] Sachs, A.B., Davis, R.W. and Kornberg, R.D. (1987) Mol. Cell. Biol. 7, 3268–3276.

[4] Query, C.C., Bentley, R.C. and Keene, J.D. (1989) Cell 57, 89–101.

[5] Scherly, D., Boelens, W., van Venrooij, W.J., Dathan, N.A., Hamm, J. and Mattaj, I.W. (1989) EMBO J. 8, 4163–4170.

[6] Nagai, K., Oubridge, C., Jessen, T.H., Li, J. and Evans, P.R. (1990) Nature 348, 515–520.

[7] Hoffman, D.W., Query, C.C., Golden, B.L., White, S.W. and Keene, J.D. (1991) Proc. Natl. Acad. Sci. USA 88, 2495–2499.

[8] Görlach, M., Wittekind, M., Beckman, R.A., Mueller, L. and Dreyfuss, G. (1992) EMBO J. 11, 3289–3295.

[9] Li, Y. and Sugiura, M. (1990) EMBO J. 9, 3059–3066.

[10] Ye, L., Li, Y., Fukami-Kobayashi, K., Gō, M., Konishi, T., Watanabe, A. and Sugiura, M. (1991) Nucleic Acids Res. 19, 6485–6490.

[11] Schuster, G. and Gruissem, W. (1991) EMBO J. 10, 1493–1502.

[12] Palmer, J.D. (1990) Trends Genet. 6, 115–120.

[13] Fukami-Kobayashi, K. (1993) Mol. Biol. Evol. (in press).

[14] Bandziulis, R.J., Swanson, M.S. and Dreyfuss, G. (1989) Gene. Dev. 3, 431–437.

[15] Jones, D.R., Taylor, W.R. and Thornton, J.M. (1992) CABIOS 8, 275–282.

[16] Saitou, N. and Nei, M. (1987) Mol. Biol. Evol. 4, 406–425.

[17] Felsenstein, J. (1985) Evolution 39, 783–791.

[18] Zahler, A.M., Lane, W.S., Stolk, J.A. and Roth, M.B. (1992) Gene. Dev. 6, 837–847.

[19] Burd, C.G., Swanson, M.S., Görlach, M. and Dreyfuss, G. (1989) Proc. Natl. Acad. Sci. USA 86, 9788–9792.

[20] Riva, S., Morandi, C., Tsoulfas, P., Pandolfo, M., Biamonti, G., Merrill, B., Williams, K.R., Multhaup, G., Beyreuther, K., Werr, H., Henrich, B. and Schäfer, K.P. (1986) EMBO J. 5, 2267–2273.

[21] Matunis, E.L., Matunis, M.J. and Dreyfuss, G. (1992) J. Cell Biol. 116, 257–269.

[22] Tian, M. and Maniatis, T. (1993) Cell 74, 105–114.

[23] Inoue, K., Hoshijima, K., Sakamoto, H. and Shimura, Y. (1990) Nature 344, 461–463.

[24] Doolittle, R.F. (1985) Trends Biochem. Sci. 10, 233–237.

[25] Patthy, L. (1991) Curr. Opin. Struct. Biol. 1, 351–361.

[26] Doolittle, R.F. (1992) Protein Sci. 1, 191–200.

[27] Nietfeld, W., Mentzel, H. and Pieler, T. (1990) EMBO J. 9, 3699–3705.

[28] Burd, C.G., Matunis, E.L. and Dreyfuss, G. (1991) Mol. Cell. Biol. 11, 3419–3424.

[29] Ge, H., Zuo, P. and Manley, J.L. (1991) Cell 66, 373–382.

[30] Krainer, A.R., Mayeda, A., Kozak, D. and Binns, G. (1991) Cell 66, 383–394.

[31] Champlin, D.T., Frasch, M., Saumweber, H. and Lis, J.T. (1991) Gene. Dev. 5, 1611–1621.

[32] Roth, M.B., Zahler, A.M. and Stolk, J.A. (1991) J. Cell Biol. 115, 587–596.

[33] Ayene, M., Preuss, U., Köhler, G. and Nielsen, P.J. (1991) Nucleic Acid Res. 19, 1273–1278.

[34] Fu, X.-D. and Maniatis, T. (1992) Science 256, 535–538.

[35] Vellard, M., Sureau, A., Soret, J., Martinerie, C. and Perbal, B. (1992) Proc. Natl. Acad. Sci. USA 89, 2511–2515.

[36] Cobianchi, F., SenGupta, D.N., Zmudzka, B.Z. and Wilson, S.H. (1986) J. Biol. Chem. 261, 3536–3543.

[37] Kay, B.K., Sawhney, R.K. and Wilson, S.H. (1990) Proc. Natl. Acad. Sci. USA 87, 1367–1371.

[38] Ball, E.E., Rehm, E.J. and Goodman, C.S. (1991) Nucleic Acids Res. 19, 397.

[39] Haynes, S.R., Rebbert, M.L., Mozer, B.A., Forquignon, F. and Dawid, I.B. (1987) Proc. Natl. Acad. Sci. USA 84, 1819–1823.

[40] Buvoli, M., Biamonti, G., Tsoulfas, P., Bassi, M.T., Ghetti, A., Riva, S. and Morandi, C. (1988) Nucleic Acids Res. 16, 3751–3770.

[41] Bell, L.R., Maine, E.M., Schedl, P. and Cline, T.W. (1988) Cell 55, 1037–1046.

[42] Grange, T., Martins de Sa, C., Oddos, J. and Pictet, R. (1987) Nucleic Acids Res. 15, 4771–4787.

[43] Lefrère, V., Vincent, A. and Amalric, F. (1990) Gene 96, 219–225.

[44] Sachs, A.B., Bond, M.W. and Kornberg, R.D. (1986) Cell 45, 827–835.

[45] Adam, S.A., Nakagawa, T., Swanson, M.S., Woodruff, T.K. and Dreyfuss, G. (1986) Mol. Cell. Biol. 6, 2932–2943.

[46] Gómez, J., Sánchez-Martínez, D., Stiefel, V., Rigau, J., Puigdomènech, P. and Pagès, M. (1988) Nature 334, 262–264.

[47] Crétin, C. and Puigdomènech, P. (1990) Plant Mol. Biol. 15, 783–785.

[48] Merrill, B.M., Lopresti, M.B., Stone, K.L. and Williams, K.R. (1987) Int. J. Pep. Res. 29, 21–39.

[49] Theissen, H., Etzerodt, M., Reuter, R., Schneider, C., Lottspeich, F., Argos, P., Lührmann, R. and Philipson, L. (1986) EMBO J. 5, 3209–3217.

[50] Spritz, R.A., Strunk, K., Surowy, C.S., Hoch, S.O., Barton, D.E. and Francke, U. (1987) Nucleic Acids Res. 15, 10373–10391.

[51] Hornig, H., Fischer, U., Costas, M., Rauh, A. and Lührmann, R. (1989) Eur. J. Biochem. 182, 45–50.

[52] Etzerodt, M., Vignali, R., Ciliberto, G., Scherly, D., Mattaj, I.W. and Philipson, L. (1988) EMBO J. 7, 4311–4321.

[53] Mancebo, R., Lo, P.C.H. and Mount, S.M. (1990) Mol. Cell. Biol. 10, 2492–2502.

[54] Amrein, H., Gorman, M. and Nöthiger, R. (1988) Cell 55, 1025–1035.

[55] Bourbon, H.-M., Lapeyre, B. and Amalric, F. (1988) J. Mol. Biol. 200, 627–638.

[56] Lapeyre, B., Bourbon, H., Amalric, F. (1987) Proc. Natl. Acad. Sci. USA 84, 1472–1476.

[57] Bourbon, H.-M. and Amalric, F. (1990) Gene 88, 187–196.