Primary structure differences between proteins C1 and C2 of HeLa 40S nuclear ribonucleoprotein particles

Barbara M.Merrill, Stanley F.Barnett, Wallace M.LeStourgeon[1] and Kenneth R.Williams

Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06510 and [1]Department of Molecular Biology, Vanderbilt University, Nashville, TN 37235, USA

Received September 22, 1989; Revised and Accepted October 11, 1989

ABSTRACT

Partial acid cleavage, comparative HPLC tryptic peptide mapping and amino acid sequencing of the C1 and C2 proteins of HeLa heterogeneous nuclear ribonucleoprotein (hnRNP) particles demonstrate that proteins C1 and C2 differ in primary structure by the presence of a 13 amino acid insert sequence in C2. This C2 insert sequence occurs after either glycine 106 or serine 107 in C1. The additional 13 amino acids that are present in C2 account for the observed molecular weight difference between the C1 and C2 hnRNP proteins on SDS polyacrylamide gel electrophoresis. Because C1 and C2 appear identical except for the 13 residue insert and because the 3′ and 5′ untranslated regions of the corresponding mRNAs also appear to be the same (Swanson *et al.*, Mol. Cell. Biol. *7*: 1731−1739), it is possible that both polypeptides are produced from a single transcription unit through an alternative splicing mechanism.

INTRODUCTION

Newly transcribed RNA quickly associates in the nucleus with multiple copies of a discrete group of proteins to form heterogeneous nuclear ribonucleoprotein (hnRNP) particles (for reviews see 1−3 and W.M. LeStourgeon, S.F. Barnett, and S.J. Northington, in *The Eucaryotic Nucleus*, Telford Press, in press). The six major proteins of core 40S hnRNP particles have apparent molecular weights on SDS polyacrylamide gel electrophoresis (PAGE) that range from 34,000 to 44,000 and, in order of increasing apparent molecular weight, have been called the A1, A2, B1, B2, C1, and C2 proteins (4). Proteins C1 and C2 are distinct from the A and B species in that the type C proteins are acidic rather than basic, they photocross-link more readily *in vivo* and *in vitro* to nucleic acids, their dissociation from RNA requires higher NaCl concentrations, they do not contain dimethyl arginine, and they are known to be highly phosphorylated (2−9). Since monoclonal antibodies against C1 and C2 inhibit RNA splicing *in vitro*, these two proteins also appear to be involved in splicing (10).

Although the C1 and C2 proteins migrate on SDS-PAGE with apparent molecular weights of approximately 42,000 and 44,000 respectively, DNA sequencing studies on a cDNA clone corresponding to the type C hnRNP proteins predict a molecular weight of only 31,931 (11). Since *in vitro* translation of a hybrid selected, 1.9 kb HeLa mRNA resulted in the production of both C1 and C2 it was proposed that both of these proteins were translated from the same mRNA (11). The 2,000 dalton difference in the apparent molecular weights of these two proteins was proposed to result from post-translational modification(s) (11). More recently, Preugschat and Wold described the isolation and characterization of a *Xenopus laevis* C protein cDNA which generated 2 proteins following *in vitro* transcription

and translation. These authors suggested alternative translation start sites to explain their result (12).

We have used partial acid cleavage to show that the cause of the apparent difference in molecular weight between HeLa C1 and C2 is not the result of alternative translation start sites. HPLC tryptic peptide mapping followed by amino acid sequencing of selected peptides demonstrate that there are internal primary structure differences between the C1 and C2 hnRNP proteins.

MATERIALS AND METHODS

Type C hnRNP protein tetramers ($C1_3C2$) were purified from 8.0×10^9 HeLa cells following procedures described elsewhere (13,14). Proteins C1 and C2 were separated from 420 μg of purified tetramer by preparative SDS-PAGE using two $0.75\times138\times95$ mm slab gels (15). The gels were stained for 10 min with 0.1% Coomassie Blue in water and destained in water until the bands were visible (30 min). The bands were excised, sliced into 1 cm lengths and soaked in electroelution buffer (25 mM Tris-HCl, 192 mM glycine, 0.1% SDS, pH 8.3) for 2 hr. The gel pieces and buffer were placed in an Elutrap (Schleicher & Schuell, Inc.) and the protein electroeluted at 200 V for 6 hr. The electroeluted protein was dialyzed against 1 liter of 0.01% SDS for 14 hr, ethanol precipitated, washed once with 70% ethanol, once with acetone, and dried under vacuum.

Previously described procedures were used to cleave proteins C1 and C2 at aspartyl-proline bonds (16,17). Briefly, 50 μl of C protein was added at a concentration of 20 μg/ml to 950 μl 75% formic acid in 7 M guanidine-HCl (Formic acid, 91.9%, Fisher Scientific Co., was diluted with 7M guanidine-HCl, Ultra Pure, Schwarz/Mann). A second 50 μl aliquot of purified C protein was added to 950 μl distilled H_2O to serve as a control. Both samples were then incubated for 22 h at 37°C. Peptides were precipitated by adding 100% trichloroacetic acid to a final concentration of 15% and incubating 1 h at 0°C. The precipitate was collected by centrifugation, washed once with 70% ethanol, dried under vacuum, dissolved in electrophoresis sample buffer and resolved in an 8.75% gel as described above. The gel was then silver stained.

In order to isolate tryptic peptides from C1 and C2 the dried, electroeluted proteins were dialyzed versus 0.10% SDS and acetone precipitated. After dissolving in 8 M urea, 500 pmol aliquots of each, as determined by amino acid analysis, were digested with trypsin as described previously (18). The resulting peptides were applied onto a Vydac C-18 column (4.6×250 mm) that had been equilibrated in 0.05% trifluoroacetic acid, 1.6% CH_3CN and that was then eluted with increasing concentrations of buffer B (0.05% trifluoroacetic acid, 80% acetonitrile) as follows: 0−63 min (2.37−5% B), 63−95 min (37.5−75% B), 95−105 min (75−98% B). Selected peptides from each digest were applied directly onto an Applied Biosystems Model 470A Protein Sequencer that was connected to a Model 120A Phenylthiohydantoin Amino Acid Analyzer.

RESULTS

One possible explanation for the apparent difference in the molecular weights of C1 and C2 could be the utilization of two different start sites on a single mRNA. To rule out this possibility, both proteins were cleaved at the single aspartyl-proline bond that, based on cDNA sequencing studies, is close to the NH_2-terminus of these proteins (11). As shown in Fig. 1, incubation of the purified C proteins with 75% formic acid at 37°C, conditions that are relatively specific for hydrolysis of aspartyl-proline bonds, results in

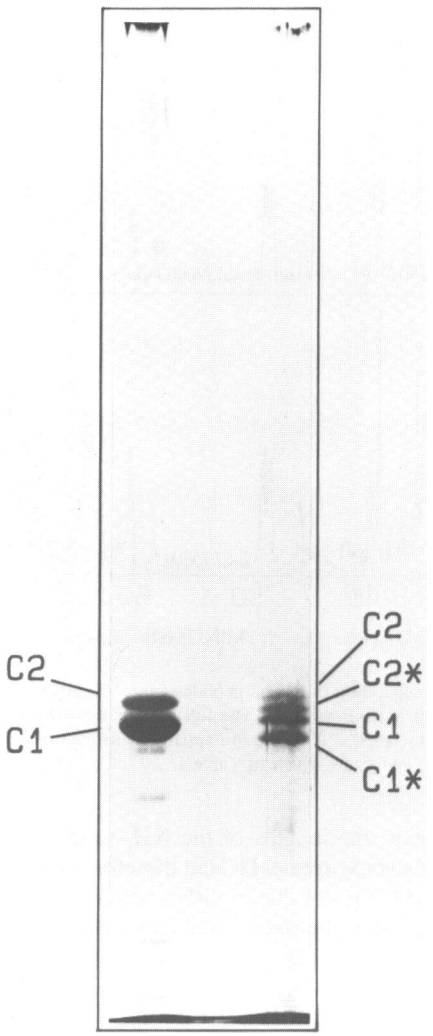

Fig. 1. SDS polyacrylamide gel electrophoresis before (left lane) and after (right lane) partial acid cleavage of the C1 and C2 hnRNP proteins. The bands labelled C1* and C2* are the high molecular weight products of the cleavage reaction.

55–60% cleavage of both C1 and C2. In each case a new band is produced, C1* and C2* respectively, which migrates on SDS-PAGE just ahead of the parent protein. The decrease in apparent molecular weight (about 1000) is consistent with the shift expected for the removal of the first ten amino acids of each protein by cleavage between aspartic acid-10 and proline-11 (The resulting amino-terminal fragment would run at the ion front of the gel). If the reason for the observed molecular weight difference between C1 and C2 was in this NH_2-terminus region, only one new resolvable band would be generated. Since two new polypeptides are generated (C1* and C2*) and their migration rates are

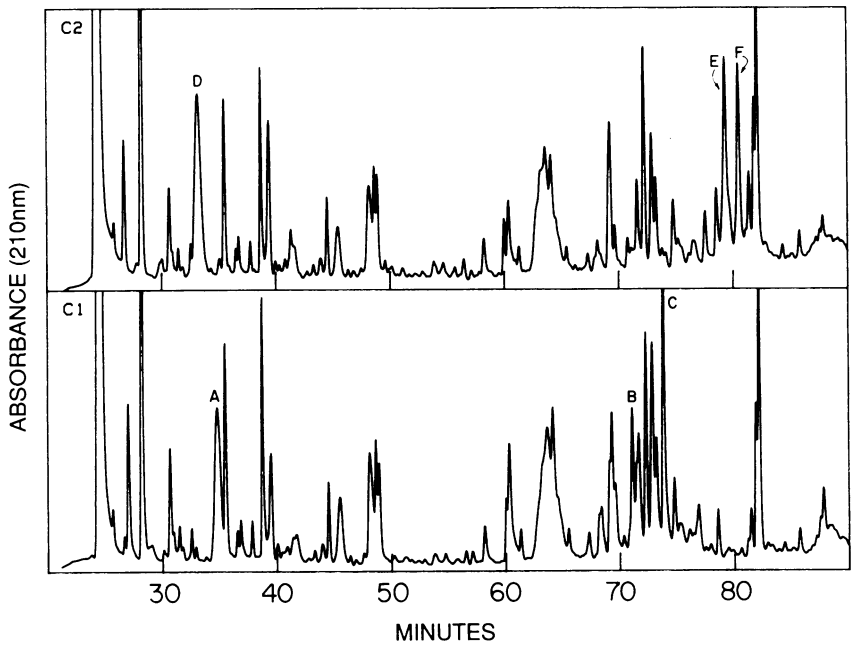

Fig. 2. Reversed-phase HPLC separation of tryptic peptides from 500 pmol of the C1 (bottom chromatogram) and the C2 (top chromatogram) hnRNP proteins. The full scale absorbance was 0.075 for the C1 and 0.060 for the C2 chromatogram. Peaks labelled A and D are apparently artifact peaks. The amino acid sequences of the peptides in peaks B, C, E, and F are given in Table I.

shifted by a similar amount, the lengths of the NH_2-termini of the C1 and C2 proteins appear to be the same at least to proline-11, and therefore, the observed molecular weight difference between C1 and C2 is not due to differences in their NH_2-termini. This appears to rule out the hypothesis that alternative translation start sites are responsible for the differences between C1 and C2 (12).

To identify structural differences between C1 and C2, both proteins were digested with trypsin and then subjected to comparative HPLC peptide mapping. As shown in Fig. 2 there are three absorbance peaks in the C1 chromatogram, labelled A, B, and C, that do not appear to align with corresponding peaks in the C2 chromatogram. Similarly, there are three absorbance peaks in the C2 chromatogram, labelled D, E, and F that do not align with corresponding peaks in the C1 chromatogram. Peaks A and D appear to be artifact peaks since neither was present in an otherwise identical C1 chromatogram that was run (data not shown) and in addition, both of these peaks failed to sequence. The amino acid sequences shown in Table I for the C1 peptides labelled B and C in Fig. 2 agree exactly with the amino acid sequences for residues 100−116 and 100−121 respectively as predicted from DNA sequencing studies on a cDNA clone corresponding to a human C-type hnRNP protein (11). Peptide C, which probably ends at arginine 122, appears to result from incomplete cleavage at the arginine-aspartic acid sequence at residues 117−118. Although the first 8 residues of sequence shown in Table I for the C2 peptides

Table 1. Amino acid sequences of tryptic peptides that contain primary structure differences in C1 and C2[a]

Protein	Peptide[a]	Amino Acid Sequence[b]
C1	B	Ser-Ala-Ala-Glu-Met-Tyr-Gly-Ser-Ser-Phe-Asp-(Leu)-(Asp)-(Tyr)-(X)-(Phe)-(Gln)[c]
C1	C	Phe-Asp-Leu-Asp-Tyr-Asp-Phe-Gln-Arg-Asp-Tyr-Tyr-Asp[d]
C2	E	Ser-Ala-Ala-Glu-Met-Tyr-Gly-Ser-Val-Thr-Glu-His-Pro-Ser-Pro-Ser-Pro-Leu-Leu-Ser-Ser-Ser-Phe-Asp-Leu-Asp-Tyr-(X)-Phe-Gln-(Arg)-(X)-(Tyr)
C2	F	Ser-Ala-Ala-Glu-Met-Tyr-Gly-Ser-Val-Thr-Glu-His-Pro-Ser-Pro-(Ser)-(X)-Leu-Leu-(Ser)-(Ser)-(X)-Phe-(Asp)

[a] Peptide designations correspond with the labelled absorbance peaks in Fig. 2.
[b] Tentative identifications are in parentheses.
[c] Corresponds to residues 100–116 as predicted from the cDNA sequence (19).
[d] Corresponds to residues 100–121 as predicted from the cDNA sequence (19).

labelled E and F in Fig. 2 correspond to residues 100–107 as predicted from the cDNA sequence (11), the next 13 amino acids are not predicted from the cDNA sequence. This 13 amino acid sequence, either Ser-Val-Thr-Glu-His-Pro-Ser-Pro-Ser-Pro-Leu-Leu-Ser- or Val-Thr-Glu-His-Pro-Ser-Pro-Ser-Pro-Leu-Leu-Ser-Ser, represents in the former case an insertion after glycine 106 and in the latter case an insertion after serine 107. Since both C2 peptides labelled 'E' and 'F' in Fig. 2 and Table I clearly contain the same 13 amino acid insert sequence, the reason for the differing elution positions of these two peptides is not clear. One possible explanation for observing both peptides 'E' and 'F' could be that one of these peptides contains a post-translational modification that is difficult to detect by amino acid sequencing. Another possible explanation is that they represent different overlapping peptides.

The predicted molecular weight for the C2 insert is 1,332 which accounts for most of the apparent molecular weight difference between C1 and C2 hnRNP proteins on SDS-PAGE. The C1 protein has a mobility on SDS-PAGE that is much less than expected based on its predicted amino acid sequence. Although the sequence-predicted molecular weight of C1 is 31,931 (11), the mobility of C1 on SDS-PAGE actually corresponds to a 42,000 dalton protein. This discrepancy in molecular mass has previously been ascribed to the asymmetric charge distribution in the C proteins, particularly the very acidic carboxy-terminal domain, which may bind SDS poorly (11). If the assumption is made that C2 also migrates slower than expected on SDS gel electrophoresis, then C2 would be predicted to have an actual molecular weight of $(31{,}931/42{,}000) \times 44{,}000$, where 44,000 daltons is the apparent molecular weight of C2 on SDS-PAGE (4). The resulting value of 33,450 for C2 corresponds closely with the sequence predicted value of 33,263. While it is possible that there are other small primary structure differences between C1 and C2, the lack of any other significant differences in the comparative HPLC tryptic peptide maps for C1 and C2 and the close agreement between the predicted and actual molecular weight for C2 suggests that the 13 amino acid insert sequence represents the only difference in primary structure between C1 and C2.

DISCUSSION

The presence of 13 additional amino acids in C2 means that C1 and C2 are not translated from the same mRNA, as previously thought (11). Rather, the cDNA clone isolated and sequenced by Swanson *et al.* (11) would correspond to the mRNA for C1. The mRNAs for C1 and C2 must be very similar if not identical in sequence in both the 3′ and 5′-untranslated regions, since both hybridize to probes based on the C1 cDNA sequence in these regions (11). The similarity between the C1 and C2 mRNAs also accounts for the hybrid selection of the C2 mRNA using a cDNA clone for C1 and explains why the hybrid selected mRNA generates 2 proteins in two different *in vitro* translation systems (11). It is not known if C1 and C2 differ in their posttranslational modifications as suggested by Swanson *et al.* (11). However, it is not necessary to invoke variations in post-translational modification to explain the molecular weight difference between C1 and C2. Most, if not all, of this difference can be accounted for by the 13 amino acid insert present in C2.

Preugshat and Wold (12) have recently determined the sequence of a cDNA for the *Xenopus laevis* C protein. These authors also concluded that the C1 and C2 proteins are generated from a single mRNA. Two different cDNAs, varying only in the length of their 5′ untranslated leaders, were described. Following *in vitro* transcription and translation, one clone generated two proteins and the other only one. These authors suggested that the length of the 5′ leader influences the use of alternative translation start sites. It is not known if the C protein gene structure in *Xenopus laevis* and *Homo sapiens* is conserved, but we have demonstrated that the generation of human C1 and C2 proteins does not result from the use of alternative translation start sites. Because the identity of the higher molecular weight band found by Preugschat and Wold (12) was not conclusively demonstrated to be C2 and the ratio of the amounts of the putative C1 and C2 proteins was quite variable, it is possible that the higher molecular weight protein was an artifact of *in vitro* transcription or translation.

One reasonable explanation for the 13 residue insert in C2 could be the use of an alternative splice site. This would be consistent with the otherwise identity between the two proteins. Alternative splicing is a common way of generating alternative protein isoforms (19). Biamonti *et al.* (20) have recently found evidence that the gene for another core hnRNP protein, A1, can produce two different mRNAs derived by alternative splicing of the transcript. Because of the multiple serine residues near the insertion site and the additional serine residues at the end of the insert, it is not possible to locate the exact position where the C2 insert begins. It is interesting to note, however, that one of the two possible start sites for the insert is a valine residue, which is encoded for by GUX, a recognition signal for the 5′-end of a splice junction (reviewed in reference 21). *In vitro* translation of C protein mRNA (11) yields C1 and C2 in the 3:1 molar ratio found in monoparticles (4) and in the C protein tetramer (22). If alternative splicing occurs, it appears that the short exon encoding the 13 residue insert is used only once in four splicing events.

One function of the 13 amino acid insert present in C2 is to modulate the protein/protein interactions responsible for stable C protein complex formation. The type C proteins of HeLa 40S nuclear ribonucleoprotein particles copurify under native conditions as a stable complex that corresponds to an anisotropic tetramer with an estimated molecular weight of 135,500 (22). This complex contains three C1 and a single C2 monomer and is not dissociated by 0.5% sodium deoxycholate or NaCl concentrations in the range of 0.09−2.0 M. The molar ratio of proteins in 40S hnRNP particles is constant (3A1, 3A2, 1B1, 1B2, 3C1, 1C2). Chemical cross-linking of native hnRNP particles generates homotypic trimers

of C1, and highly purified native C protein tetramers ($C1_3C2$) participate in hnRNP particle reconstitution. Thus the tetrameric C protein complex probably represents a basic structural unit in these RNA packaging complexes (22). The ability of C1 and C2 to form tetramers of fixed stoichiometry ($C1_3C2$) must be either a direct or indirect result of the 13 amino acid insert sequence.

The position of the 13 amino acid insert in C2 is interesting in that it occurs just after a serine residue that represents a potential phosphorylation site. The proteins in 40S hnRNP particles are known to be phosphorylated by at least two different enzymes (23–25). One of these is a casein kinase type II enzyme that has been shown to phosphorylate type C hnRNP proteins (5,26). Studies on synthetic peptide analogues demonstrate that the best substrates for this enzyme contain serine residues that are followed by clusters of acidic amino acids (27,28). Of particular importance are acidic residues at postions +3 and +5 relative to the phosphorylated serine. Based on these criteria C1 contains two potential casein kinase type II sites at serine residues 107 and 247 (Fig. 3). Although the cluster of serine residues spanning positions 225–228 in C1 has been suggested previously to represent a potential casein kinase type II site (11), this site appears less likely in that it is followed by the basic sequence valine-lysine-lysine, and thus lacks the adjacent acidic residues that appear to be an important determinant of the specificity of this enzyme. Although serine 220 is also followed by acidic residues, this site appears less likely because it is immediately preceded by a lysine. Basic residues that are immediately prior to serine residues have been shown to significantly decrease phosphorylation by casein kinase type II at those sites (29). Since the potential phosphorylation site at serine 107 in C1 occurs at the position of the insertion of thirteen amino acids in protein C2, this site is effectively moved to serine 120 in C2. It is possible that changing the position of this potential

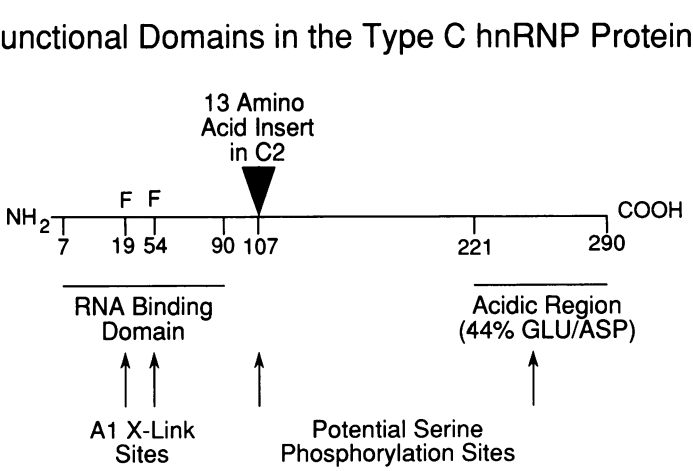

Fig. 3. Functional and structural domains in the type C hnRNP proteins. The C1 hnRNP protein (as pictured) contains 290 amino acids (19) and has a putative RNA binding domain that spans residues 7–90. Phenylalanine residues 19 and 54 are predicted to be at the oligonucleotide:type C hnRNP protein interface based on sequence homologies with A1 (19). The 13 amino acid insert in C2 occurs either after glycine 106 or serine 107. There are two putative phosphorylation sites at serine residues 107 and 247. The COOH-terminal domain (residues 221–290) is unusually acidic in that it contains 44% glutamic and aspartic acid.

phosphorylation site could modulate its use *in vivo* and that if such a posttranslational modification occurs it might effect the functional properties of the C2 protein as well as the $C1_3C2$ tetramer.

The C2 insert site is of further interest because it is adjacent to a region of the type C proteins (approximately residues 7–90) that, as depicted in Fig. 3, probably represents an RNA-binding domain. This region of the C protein contains sequence homologies with regions of numerous other RNA-binding proteins, including the A1 and A2 hnRNP proteins (for a review, see B.M. Merrill and K.R. Williams in *The Eucaryotic Nucleus*, Telford Press, in press). By analogy with the A1 hnRNP, two of the most highly conserved residues in this region, phenylalanine residues 19 and 54 in the C1 sequence (11) may be directly involved in RNA-binding. The homologous phenylalanine residues in the A1 hnRNP protein are the only sites of covalent attachment when an oligonucleotide is photocrosslinked to A1 (18).

While most of the structure/function relationships that have been postulated for the C proteins are still speculative, it appears certain that the thirteen amino acid insert sequence in C2 accounts for the ability of this protein to form an anisotropic tetrameter of fixed stoichiometry with C1. As more information concerning the structure and function of the C protein tetramer becomes available, it will become possible to evaluate the role of this complex in RNA packaging and pre-mRNA maturation.

ACKNOWLEDGEMENTS

This work was supported by Public Health Service grant GM31539 to KRW and by NSF grant DCB85-12035 to WML. KRW is supported by the Howard Hughes Medical Institute.

REFERENCES

1. Chung, S.Y., and J. Wooley. (1986) Proteins **1**:195–210.
2. Conway, G., J. Wooley, T. Bibring, and W.M. LeStourgeon. (1988) Mol. Cell. Biol. **8**:2884–2895.
3. Dreyfuss, G. (1986) Ann. Rev. Cell Biol. **2**:457–495.
4. Beyer, A.L., M.E. Christensen, B.W. Walker, and W.M. LeStourgeon. (1977) Cell **11**:127–138.
5. Holcomb, E.R., and D.L. Friedman. (1984) J. Biol. Chem. **259**:31–40.
6. Mayrand, S., B. Setyono, J.R. Greenberg, and T. Pederson. (1981) J. Cell Biol. **90**:380–384.
7. Schweiger, A., and G. Kostka. (1985) Biochim. Biophys. Acta. **826**:87–94.
8. van Eekelen, C.A., T. Riemen, and W.J. van Venrooij. (1981) **130**:223–226.
9. Wilk, H., H. Werr, D. Friedrich, H.H. Kiltz, and K.P. Schafer. (1985) Eur. J. Biochem. **146**:71–81.
10. Choi, Y.D., P.J. Grabowski, P.A. Sharp, and G. Dreyfuss. (1986) Science **231**:1534–1539.
11. Swanson, M.S., T.Y. Nakagawa, K. LeVan, and G. Dreyfuss. (1987) Mol. Cell Biol. **7**:1731–1739.
12. Preugschat, F. and B. Wold. (1988) Proc. Natl. Acad. Sci. USA. **85**:9669–9673.
13. Barnett, S.F., W.M. LeStourgeon, and D.L. Friedman. (1988) J. Biochem. Biophys. Meth. **16**:87–98.
14. Barnett, S.F., S. Northington, and W. LeStourgeon. (1988) *In* J. Dahlberg and J.N. Abelson, (eds.) Methods in Enzymology, Academic Press, New York.
15. Laemmli, U.K. (1970) Nature **227**:680–685.
16. Jauregui-Adell, L., and J. Marti. (1975) Analytical Biochemistry. **69**:468–473.
17. Landon, M. (1977) , pp. 145–149. *In* C.W. Hirs and S.N. Timasheff, (eds), Methods in Enzymology, Academic Press, New York.
18. Merrill, B.M., Stone, K.L., Cobianchi, F., Wilson, S.H., and K.R. Williams. (1988) J. JBiol. Chem. **263**:3307–3313.
19. Breitbart, R.E., A. Andreadis, and B. Nadal-Ginard. (1987) Ann. Rev. Biochem. **56**:467–495.
20. Biamonti, G., Buvoli, M., Bassi, M.T., Morandi, C., Cobianchi, F., and Riva, S. (1989) J. Mol. Biol. **207**:491–503.
21. Sharp, P.A. (1987) Science **235**:766–771.
22. Barnett, S.F., D.L. Friedman, and W.M. LeStourgeon (1989) Mol. Cell. Biol. **9**:492–498.
23. McGregor, C.W., and J.T. Knowler. (1987) Mol. Biol. Reports **12**:85–92.

24. Periasamy, M., C. Brunel, and P. Jeanteur. (1979) Biochimie. **61**:823–826.
25. Wilks, A.F., and J.T. Knowler. (1981) Biochim. Biophys. Acta **652**:228–233.
26. Friedman, D.L., N.J. Kleinman, and F.E. Campbell. (1985) Biochim. et Biophys. Acta **847**:165–176.
27. Hathaway, G.M., and J.A. Traugh. (1982) Current Topics in Cellular Regulation. **21**:101–127.
28. Marin, O., F. Meggio, F. Marchiori, G. Borin, and L.A. Pinna. (1986) Eur. J. Biochem. **160**:239–244.
29. Meggio, F., F. Marchiori, G. Borin, G. Chessa, and L.A. Pinna. (1984) J. Biol. Chem. **259**:14576–14579.

This article, submitted on disc, has been automatically converted into this typeset format by the publisher.