466–473 Nucleic Acids Research, 1997, Vol. 25, No. 3

Molecular weight abnormalities of the CTCF transcription factor: CTCF migrates aberrantly in SDS–PAGE and the size of the expressed protein is affected by the UTRs and sequences within the coding region of the CTCF gene

Elena M. Klenova*, Robert H. Nicolas[1,+], Sally U[1], Alexander F. Carne[1], Robin E. Lee, Victor V. Lobanenkov[2] and Graham H. Goodwin[1]

Genetics Unit, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK, [1]Institute of Cancer Research, Chester Beatty Laboratories, 237 Fulham Road, London SW3 6JB, UK and [2]Fred Hutchinson Cancer Research Center, 1124 Columbia Street, Seattle, WA 98104, USA

Received November 21, 1996; Accepted December 9, 1996

ABSTRACT

CTCF belongs to the Zn finger transcription factors family and binds to the promoter region of c-*myc*. CTCF is highly conserved between species, ubiquitous and localised in nuclei. The endogenous CTCF migrates as a 130 kDa (CTCF-130) protein on SDS–PAGE, however, the open reading frame (ORF) of the CTCF cDNA encodes only a 82 kDa protein (CTCF-82). In the present study we investigate this phenomenon and show with mass-spectra analysis that this occurs due to aberrant mobility of the CTCF protein. Another paradox is that our original cDNA, composed of the ORF and 3′-untranslated region (3′-UTR), produces a protein with the apparent molecular weight of 70 kDa (CTCF-70). This paradox has been found to be an effect of the UTRs and sequences within the coding region of the CTCF gene resulting in C-terminal truncation of CTCF-130. The potential attenuator has been identified and point-mutated. This restored the electrophoretic mobility of the CTCF protein to 130 kDa. CTCF-70, the aberrantly migrating CTCF N-terminus *per se*, is also detected in some cell types and therefore may have some biological implications. In particular, CTCF-70 interferes with CTCF-130 normal function, enhancing transactivation induced by CTCF-130 in COS6 cells. The mechanism of CTCF-70 action and other possible functions of CTCF-70 are discussed.

INTRODUCTION

CTCF is a novel 11 Zn-finger transcription factor, present in nuclear extracts as a protein with an apparent MW of 130 kDa (CTCF-130) on SDS–PAGE. It was discovered for its ability to bind to an unusually long 45 bp GC-rich sequence containing three regularly spaced repeats of the core sequence CCCTC within the chicken *c-myc* promoter (1). However, its ability to recognise diverged sequences by employing different combinations of zinc fingers has recently been demonstrated (2). Since deregulation of the *c-myc* proto-oncogene is causally related to oncogenesis (3,4), identification and characterisation of factors regulating its transcription may provide insight into molecular mechanisms of normal and aberrant *c-myc* expression (3–6). Thus, CTCF features were further investigated (1,7) and its cDNA cloned (8).

CTCF has been found to be expressed in different chicken tissues in multiple forms (130, 97, 80, 73, 70 and 55 kDa) as detected by anti-CTCF antibodies (8). In various cell types the CTCF protein is encoded in a 4.1 kb mRNA, with the longest open reading frame (ORF) (2184 bp) of the cDNA predicting a 728 amino acid protein of 82 kDa. This contradicts the observed MW of 130 kDa. The discrepancy between the theoretical (82 kDa) and apparent (130 kDa) MWs of the CTCF protein products could be due to a number of possibilities: (i) there is an additional exon in the primary transcript, which is missing in our cDNA possibly generated by alternative splicing (9–12); (ii) post-translational modifications (13–19) could change the mobility of the CTCF protein; (iii) particular amino acid composition (20–23) could lead to CTCF anomalous electrophoretic migration. These possibilities are discussed in this paper.

Of the other multiple forms of CTCF expressed, in this study we investigated the appearance of the 70 kDa protein. Previously, it was found that when the original CTCF cDNA containing the entire ORF and the 3′-untranslated region (3′-UTR) was expressed in cells, it produced the 70 kDa protein product, but not 82 or 130 kDa (8). We investigated the following options for this discrepancy: (i) post-translational processing (24–26) or protein splicing (26); (ii) transcription and (iii) translation attenuation (27–29).

In this study we show that that the discrepancy between the 82 and the 130 kDa MWs of the CTCF protein was due to anomalous migration in the SDS–PAGE (mechanism 3) and that the N- and C-terminal domains participate in this anomaly. We also show that the CTCF-70 protein represents a truncated version of the full length CTCF protein (CTCF-130), corresponding to the N-terminus, and is

* To whom correspondence should be addressed. Tel: +44 1865 275316/325; Fax: +44 1865 275318; Email: klenovae@bioch.ox.ac.uk

+Present address: Imperial Cancer Research Fund, PO Box 123, 44 Lincoln's Inn Fields, London WC2A 3PX, UK

probably generated by premature termination of translation of the CTCF mRNA at a specific site within the CTCF coding region. This depends on the sequences within the UTRs and the coding region. Silent nucleotide substitution (i.e. mutation without changing the CTCF amino acid sequence) at the potential attenuator site inside the ORF restored the size of CTCF-70 to 130 kDa.

Additionally, we demonstrate that the 70 kDa protein, corresponding to the abnormally migrating CTCF N-terminal domain, enhanced transactivation induced by the full length wild type (wt) CTCF in COS6 cells.

MATERIALS AND METHODS

Recombinant plasmids

Plasmids pEEc and pBBc were constructed by inserting the full length 3.786 kb *Eco*RI–*Eco*RI fragment from plasmid p900-911 and the 3.647 kb *Eco*RI–*Eco*RI fragment from p900 (8), respectively, into the eukaryotic expression vector pSG5, controlled by early SV40 promoter (Stratagene). pHHc was obtained by deleting a part of the 3′-UTR of pBBc with *Hin*dIII, resulting in a 2.855 kb insert (Fig. 1A). '*Pst*I truncated' versions of all three plasmids (pBBcΔPP, pHHcΔPP and pEEcΔPP) were obtained by 'in-frame' *Pst*I–*Pst*I deletion of 840 bp within the coding sequences. pBBc, pHHc, pBBcΔPP, pHHcΔPP and pEEcΔPP *myc*-tagged derivatives were obtained by introducing a 294 bp (*Hin*cII–*Bam*HI) fragment, encoding six *myc* tag epitopes, from the plasmid p6*Myc* Tags-D (30) into the *Msc*I site at the 3′-terminal end of the CTCF ORF (Fig. 1A). The pBBc-mut plasmid, containing point-mutations within the region of a proposed attenuator, was generated by PCR using two oligonucleotides beginning at nucleotide 385: 5′-TGATGATGGAGCACCTGGAT-3′ and nucleotide 836: 5′-GAAGCTTACTCTTCTTGTTCTTCTTGGTCTTCTTG-3′ (Fig. 1B). Nucleotide numbers correspond to the 5′-end of the longest CTCF cDNA (8). Mutated bases are shown underlined. The resulting mutated plasmid pBBc-mut was verified by sequencing. The reporter plasmid p90TKLuc was obtained by ligating the dimeric CTCF binding site (synthetic 90 oligomer) from the chicken *c-myc* promoter (–219 to –180 bp) (8) into the *Hin*dIII–*Xho*I site of the pTK Luciferase vector, based on pXp2 Luciferase vector (31). N- and C-terminal domains of the CTCF protein (117–830 bp *Sma*I–*Hin*dIII fragment and 1908–3786 bp *Msc*I–*Eco*RI fragment, respectively) were subcloned in frame into pGex 2TK (Pharmacia), generating pGST-N and pGST-C. The Zn-finger domain (541–2037 *Pvu*II–*Eco*RI fragment from the p900 plasmid) was cloned into pGEX-3X (Pharmacia), generating pGST-Zn. Regions of DNA around the junctions were checked by sequencing.

Cells and transfections

COS6 cells were grown at low (40–50%) density in Dulbecco's modified Eagle's medium supplemented with 10% foetal bovine serum, and transfected with 1–10 μg of pBBc, pHHc or pEEc plasmids by the calcium phosphate precipitation method with 2× DNA precipitation buffer following the manufacturer's instructions ('5 Prime-3 Prime', Inc^{R}, CP Laboratories, UK). Cell extracts were made 48 h post-transfection. For co-transfection assays 1.25×10^5 COS6 cells were plated in 16 mm dishes and 1 μg of reporter (p90TK Luc) and 0.5–2.5 μg of expressor (pHHc or pBBc) were introduced with Lipofectamine (2.5 μl/dish) following the manufacturer's instructions (Gibco BRL). The total amount of the DNA transfected was adjusted to 3.5 μg with the 'empty' pSG5 vector. Cells were harvested 16–18 h after transfection. Luciferase assays were performed with a 'Dual-Light^{TM}' kit (Tropix, USA) following the manufacturer's instructions using a TD-20e luminometer (Turner, USA).

Western blotting

Total cell lysates were treated with DNAse I in the presence of 5 mM $MgCl_2$, electrophoresed on 10% SDS–PAGE gels, transferred onto Immobilon P membranes (Millipore, Bedford, MA) by semi-dry blotting, and probed with either the chicken-specific anti-CTCF antibody Ab2 (8) or monoclonal 9E10 anti-*myc* tag antibody (33). CTCF or *myc*-tag-fusion proteins were visualized by the standard ECL procedure with an ECL detection system (Amersham).

In vitro translation

Aliquots of 1 μg of the pBBc, pEEc and pHHc DNA were translated *in vitro* using the TNT^{R} coupled reticulocyte lysate system, according to the manufacturer's protocol (Promega) from the T7 promoter present in the pSG5 expression vector. The synthesised proteins were resolved on 10% SDS–PAGE gels, and blotted onto Immobilon-P membrane where necessary.

Protein preparation and molecular weight determination

The glutathione-S-transferase (GST)-fusions were expressed in TG-1 cells and purified on glutathione–Sepharose as previously described (32), then to homogeneity by anion exchange FPLC using a mono-Q column (Pharmacia) and eluted at ~0.2 M NaCl with a salt gradient (0.5–0.1 M) in a 20 mM Tris–HCl buffer pH 8.0. The purified GST-fusions were cleaved with thrombin, then treated sequentially with benzamidine beads (Sigma) and glutathione–Sepharose (Pharmacia) to remove the protease and the GST, respectively. The molecular weight of the cleaved proteins was determined by comparing their mobility with that of Rainbow molecular weight markers (Amersham) after electrophoresis on 10% polyacrylamide–SDS gels. The cleaved proteins were also sequenced from their N-termini in order to confirm their identities. The true molecular weights were determined by matrix-assisted laser desorption mass spectrometry (MALD-MS) using the Finnigan Lasermat 200 mass spectrometer. The N- or C-portions of the CTCF protein were isolated by FPLC on a MonoQ column (Pharmacia), the protein samples were reduced with 20 mM dithiothreitol, desalted by HPLC on a C4 reverse phase column (Brownlee) with acetonitrile 0.1% trifluoroacetic acid. After chromatography the sample was dried under vacuum, taken up in 10% methanol, and then 0.5 ml was mixed with 0.5 ml of sinopinic acid matrix (10 mg/ml) solution in 70% acetonitrile on a slide following the manufacturer's instructions. The mass spectrometer was calibrated with bovine serum albumin (Sigma, M_r = 66 430) for the uncleaved proteins and carbonic anhydrase II (Sigma, M_r = 29.023) for the protein of molecular weight <30 000 Da. Each spectrum was determined at least three times and the mean values calculated.

RESULTS

The difference between the predicted (82 kDa) and apparent (130 kDa) molecular weights of the full length CTCF protein is due to aberrant migration of the N- and C-termini

To explain the discrepancy between the predicted 82 and 130 kDa MWs of the CTCF protein, we first considered mechanism 1, i.e.,

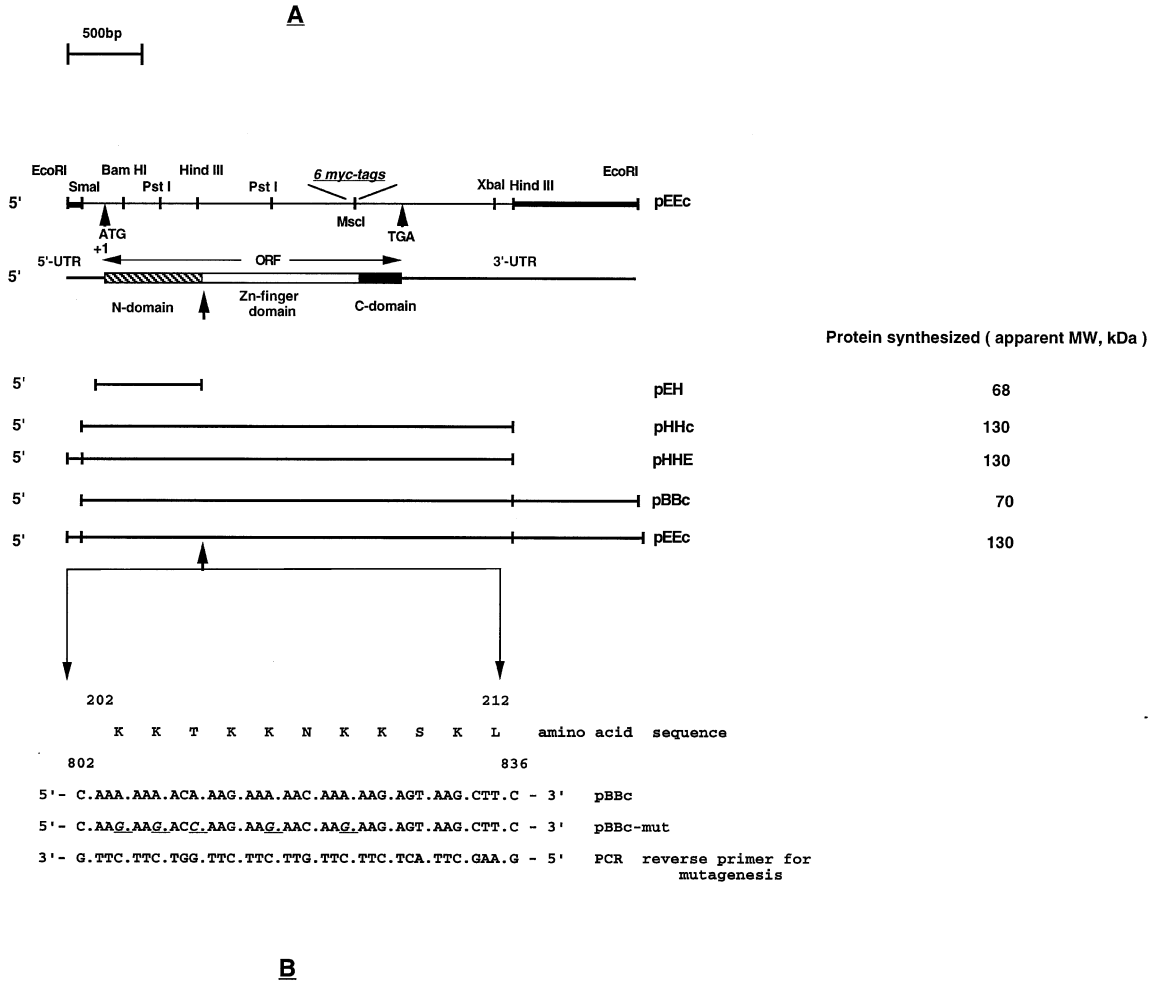

Figure 1. Maps of the the CTCF cDNA, CTCF protein, pCTCF expressing constructs and PCR primer sequences. (**A**) Restriction map of the full length CTCF cDNA and the position of the 5′- and 3′-UTRs, the N- and C-terminal domains and the Zn-finger domain of the CTCF protein. Shown below are the corresponding maps of the CTCF truncated versions, showing which parts of the cDNA have been deleted in each construct. The molecular weights of the proteins synthesised from these plasmids are indicated. (**B**) Point-mutations within the potential attenuator site (mutated bases are underlined). The mutations introduced do not change the amino acid sequence in this region, but disrupt poly(A) stretches.

the possibility of an additional exon in the primary transcript which is missing in the cDNA. All our investigations, including screening various cDNA libraries, RACE and RT–PCR on various total and polyA$^+$ mRNA preparations, failed to identify any additional coding sequences which could produce a 130 kDa form of CTCF protein. Therefore mechanism 1 was ruled out and the remaining options examined: mechanism 2, i.e. post-translational modifications of the protein and mechanism 3, i.e. CTCF-82 anomalous electrophoretic migration. In order to investigate these options, three parts of the CTCF cDNA (encoding the N-terminus, Zn-finger domain and C-terminus) were subcloned into pGex2TK vector (N- and C-termini) or pGex3T (Zn-finger domain). Recombinant GST-fusion proteins were expressed in *Escherichia coli*, purified and analysed by mass-spectrometry. Mass-spectrometry analysis did not reveal any post-translational modifications (mechanism 2). As shown in Table 1, the experimental molecular weights of the GST-N fusion and N- and C-termini alone, determined by MALDI-MS, correspond well with their theoretical MWs. However, SDS–PAGE analysis demonstrated anomalous electrophoretic mobility of the GST-fusion proteins and the N- and C-termini. GST-Zn migrates as a 98 kDa protein in contrast to its predicted MW of 81.4 kDa (20% anomalous migration, Fig. 2A, lane 4). This fusion protein is rather unstable, giving rise to the products of degradation appearing as a smear after the staining with the anti-GST-antibodies. GST-C migrates as a 70 kDa protein, in contrast with its predicted size of 44 kDa (59% aberrant migration, Fig. 2A, lane 2). GST-N appears as a predominant band of ~97 kDa in contrast to its predicted size of 53.8 kDa (79% of aberrant migration, Fig. 2A, lane 3), with two minor bands most likely resulting from degradation. It is worth noting that the presence of the GST-Zn fusion protein of ~98, GST-N of 97 and GST-C of 70 kDa has been consistently detected in independent experiments involving staining with Coomassie and anti-GST-antibodies, whilst the positions of the minor bands varied. Following the GST removal the C- and N-termini migrate aberrantly at 35 and 70 in contrast to their predicted 18 and 28 kDa (94 and 160% anomalous migration, respectively, Fig. 2B, lanes 2 and 3). Thus the differences between the theoretical molecular weights and the apparent molecular weights of the C- and

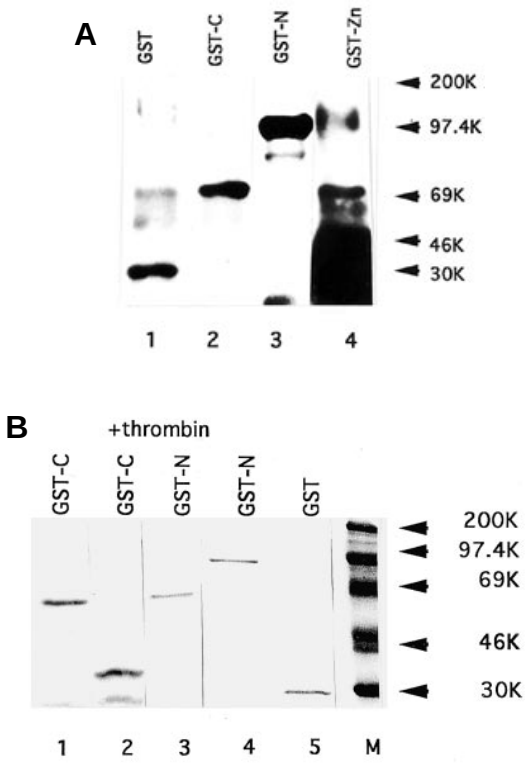

Figure 2. Analyses of the CTCF regions responsible for aberrant migration. (**A**) Three regions of the CTCF protein (N-, Zn-, C-) were cloned in frame with the GST, expressed in *E.coli*, and purified on glutathione–Sepharose beads. The eluted proteins were loaded onto a 10% SDS–PAGE gel, then blotted and probed with anti-GST rabbit polyclonal antibodies. The GST- and GST-Zn migrate at 29 kDa (lane 1) and 98 kDa (lane 4), respectively. The GST-C and GST-N migrate abnormally at 70 kDa (lane 2) and 97 kDa (lane 3), respectively. (**B**) Electrophoretic analyses of the GST-C (lane 1) and GST-N fusion proteins (lane 4) and C- and N-terminal portions of the CTCF protein after the fusion proteins were cleaved with thrombin (lanes 2 and 3, respectively). The SDS–PAGE gel was stained with Coomassie blue. Both the C-terminal and N-terminal domains migrate aberrantly at 35 and 70 kDa, respectively. The presence of the GST portion can, in some cases, be seen after the Coomassie staining (a band of ~30 kDa, lane 2), otherwise it is easily detected by Western blotting with anti-GST antibodies (data not shown).

N-terminal domains are due to anomalous electrophoretic migration of the polypeptides (mechanism 3). This would in turn account for the anomalously high apparent molecular weight of the intact CTCF protein.

The 70 and 130 kDa forms of the CTCF protein are encoded within the same ORF of the CTCF cDNA

The problem of the considerable size difference between the CTCF-130 and CTCF-70 proteins still remained to be resolved. As shown in Figure 1A, the difference between pBBc which produced the 70 kDa protein and pHHc which produced the 130 kDa protein was only in the length of the 3′-UTR (also ref. 8). To investigate this further, pBBc, pHHc and an additional construct pEEc (CTCF cDNA of 3.786 kb which differs from pBBc in having an additional 106 bp of the 5′-UTR; Fig. 1A) were expressed *in vivo* and *in vitro*. *In vitro* translation was carried out in the presence of [^{35}S]methionine. The ^{35}S-labelled protein products, made by rabbit reticulocyte lysates, and *in vivo* protein products from transfected COS6 cell lysates were electrophoresed on SDS–PAGE, blotted onto Immobilon-P membrane and autoradiographed (Fig. 3A, lanes 1–3, 4–6 respectively). The same membrane was then probed with the chicken-specific anti-CTCF antibody (8) (Fig. 3B). pEEc, which represents the full length CTCF cDNA and pHHc which is missing parts of the 3′- and 5′-UTRs produced the CTCF-130 protein in both experiments (Fig. 3A, lanes 2 and 3; Fig. 3B lanes 2 and 3, 5 and 6). However, pBBc with full length 3′-UTR but missing 5′-UTR, generates a 70 kDa protein in both *in vitro* and *in vivo* experiments (Fig. 3A, lane 1 and Fig. 3B, lanes 1 and 4). These results therefore showed that the ORF of the cloned cDNA was capable of encoding a protein that had the same electrophoretic mobility as the native protein (i.e. migrating at 130 kDa), but that a smaller protein (70 kDa) is made depending on the 5′- and 3′-UTRs of the cDNA insert.

Table 1. The molecular weight determination of CTCF protein fragments

Protein	Molecular weight/kDa			Aberrant[a]
	Theoretical	SDS–PAGE	MALDI-MS	migration (%)
GST-2TK	27.551	29	n.d.	5
cleaved	26.018	27	26.057 ± 0.046	4
GST-N	53.769	97	52.556 ± 0.023	79
cleaved[b]	27.751	70	26.594 ± 0.027	160
GST-C	44.123	70	n.d.	59
cleaved	18.105	35	18.121 ± 0.011	94
GST-Zn	81.383	98	n.d.	20
cleaved	55.418	n.d.	n.d.	n.d.

[a]The MALDI-MS molecular weight was determined and this value was used to calculate the aberrant percent migration observed in SDS–polyacrylamide electrophoresis.

[b]The molecular weight discrepancy (1157 Da) between the theoretical and that observed by MALDI-MS corresponds to the last 10 amino acid residues (1180 Da) from the C-terminus of this protein. The sequence of the last 15 residues is KTKK/NKKSKLGIHRD, with the proposed cleavage site between the lysine and asparagine residues. This sequence does not contain a known cleavage site for thrombin, so presumably the lysis took place in the bacteria or during the preparation of the protein. This conclusion is consistent with the MALDI-MS data for the uncleaved GST-N protein where the discrepancy is 1213 Da.

CTCF-70 is likely to be a product of premature termination of translation

In order to determine the origin of the 70 kDa CTCF protein, a sequence encoding six *myc*-tag epitopes was cloned in frame into the end of the CTCF Zn-finger domain of both plasmids—pBBc, producing the 70 kDa CTCF protein, and pHHc, producing the 130 kDa protein. As shown in Figure 4, after transfection into COS6 cells both plasmids generate proteins which can be recognized by the chicken-specific antibody (Fig. 4A). When the membrane was stripped of the anti-CTCF antibody and reprobed with the anti-*myc* tag antibody (9E10), only the product of pHH*c-myc* plasmid was positive, demonstrating that the p70 kDa protein is lacking the CTCF C-terminus (Fig. 4B). (The appearence of the multiple bands in the pHHc *myc*-tagged plasmid is discussed below.)

Sequences responsible for generation of the CTCF-70 and CTCF-130

Analysis of sequences within the 3′- and 5′-UTRs affecting the size of the CTCF protein products, revealed that addition of only

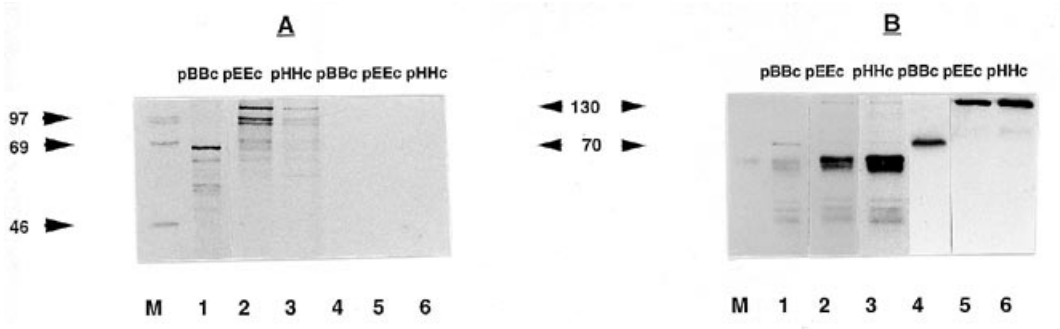

Figure 3. Analyses of pBBc, pEEc and pHHc *in vitro* and *in vivo* translation products. (**A**) pBBc, containing the ORF and the 3′-UTR of the CTCF cDNA (lane 1), pEEc containing the entire 3.7 kb CTCF cDNA (lane 2) and pHHc missing parts of the 5′- and 3′-UTRs (lane 3), were translated in rabbit reticulocyte lysates with [^{35}S]methionine, and analysed on SDS–PAGE. The pEEc and pHHc translation products migrated at 130 kDa (lanes 2 and 3), whereas the pBBc translation product migrated at ~70 kDa (lane 1). The lysates of COS6 cells transfected with pBBc, pEEc and pHHc were analysed in parallel on the same gel (lanes 4–6). The gel was blotted onto the membrane and exposed overnight. (**B**) Western blot analyses of pBBc, pEEc and pHHc *in vitro* translation (lanes 1–3) and *in vivo* (lanes 4–6) transfection products. The membrane (A) was probed with chicken-specific anti-CTCF antibodies and assayed by ECL. The protein products of 130 kDa can be seen when pEEc (lanes 2 and 5) and pHHc (lanes 3 and 6) were expressed both *in vitro* and *in vivo*. However, the protein of 70 kDa is seen in the case on pBBc expressed *in vitro* and *in vivo* (lanes 1 and 4, respectively).

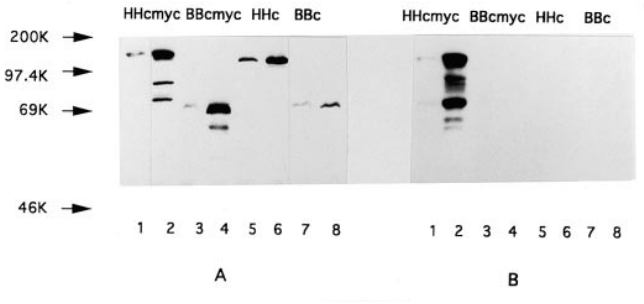

Figure 4. The 70 kDa protein is a truncated version of the 130 kDa CTCF protein representing the N-terminal domain. Total cell lysates were obtained from COS6 cells transfected with 1 or 10 µg (odd and even lanes, respectively) of CTCF expressing plasmids labelled with the *myc*-tag epitope in their C-terminal domain (lanes 1 and 2, pHH*cmyc*; lanes 3 and 4, pBBc*myc*) and the parental vectors (lanes 5 and 6, pHHc; lanes 7 and 8, pBBc). Aliquots of 20 µl of each extract were loaded onto a 10% SDS–PAGE gel, resolved and blotted onto Immobilon-P membrane. The membrane was probed with rabbit chicken specific anti-CTCF antibodies (**A**). After that the filter was stripped and reprobed with the anti-*myc* tag 9E10 mouse monoclonal antibodies (**B**). Positive staining can only be seen with lysates obtained from cells transfected with pHHc*myc* (lanes 1 and 2). The presence of multiple bands is explained in the text.

106 bp (*Eco*RI–*Sma*I fragment, Fig. 1A) of the 5′-UTR to the 3.672 kb cDNA (pBBc, Fig. 1A), restored the size of the protein expressed *in vivo* and *in vitro* to 130 kDa (pEEc, Fig. 1A). This fragment is very GC-rich (73% of GC-content, ref. 8) and its effect is specific. Thus, replacing it with other fragments (either GC- or AT-rich) did not change the size of the protein (70 kDa) produced from these hybrid cDNAs (data not shown). On the other hand, addition of a 792 bp of the 3′-UTR (*Hin*dIII–*Eco*RI fragment, pEEc, Fig. 1A) to the plasmid pHHc-130, resulting in pBBc, reduced the apparent size of the protein made from 130 to 70 kDa. However, removal of this fragment from pEEc (pHHE, Fig. 1A) did not change the size of the protein produced (130 kDa, data not shown). Thus, only particular combinations of the UTRs with sequences inside the coding region are responsible for the production of CTCF-70.

Deletion in the region containing the potential attenuator leads to production of the same size protein from pEEc, pHHc and pBBc

It has been noticed that the size of the 70 kDa protein corresponds to the size (68 kDa) of the CTCF N-terminal domain migrating abnormally (Fig. 1A, black arrow). This, together with the observation that the C-terminal domain is missing in CTCF-70 (Fig. 4), suggests the existence of a potential attenuation site within the coding region close to the end of the N-terminal domain with a likely position inside the sequence containing poly(A) stretches (Fig. 1B and ref. 8). To locate the region in pBBc where premature termination of translation occurs the in-frame *Pst*I–*Pst*I deletion, which removes the putative attenuatior site, was obtained in all three plasmids, pBBc, pHHc and pEEc (Fig. 1A). As shown in Figure 5A (lanes 1–3), after transfection in COS6 cells, all three plasmids carrying the deletion produced a protein of the same size (97 kDa). This finding suggested the presence of the potential attenuator in pBBc within this deleted region. To further prove that this deletion does prevent translation termination and the C-terminal portions of all the proteins produced are intact, we employed the same '*myc*-tag' approach as described above. For this purpose a sequence encoding six *myc*-tag epitopes was cloned in frame into the *Msc*I site (the end of the Zn-finger domain) of all three plasmids (see also Materials and Methods and Fig. 1). As shown in Figure 5A (lanes 4–6) the three plasmids after transfection in COS6 cells and Western-immunoblot with the chicken-specific CTCF antibodies, produce proteins of the same size (~110 kDa which corresponds to the size of the *Pst*I–*Pst*I deleted plasmids: 97 kDa plus the *myc*-tag component, 11 kDa). When the membrane was stripped of the anti-CTCF antibodies and reprobed with 9E10, the same bands appeared positive (Fig. 5B, lanes 4–6). The multiple bands of smaller size observed in all our *myc*-tagged proteins (including pHHc*myc*, Fig. 4A and B, lane 2) are not products of premature termination, but most likely resulted due to proteolysis of the *myc*-tagged proteins containing foreign sequences and therefore unstable because: (i) they are detected only with the 9E10 anti-*myc*-tag antibodies and not with the anti-N-terminal CTCF antibodies; (ii) they are specific only to the *myc*-tagged plasmids; and (iii) the same fragments can be seen in all three *myc*-tagged

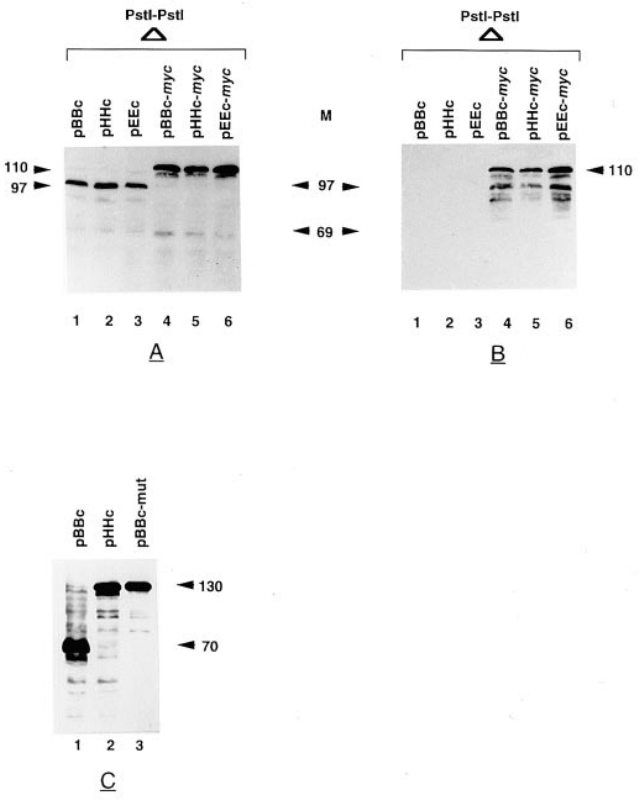

Figure 5. Deletion and silent nucleotide substitutions within the potential attenuator site restore the size of the CTCF-70 encoded in pBBc. Aliquots of 10 µg of pBBc, pHHc and pEEc with an 'in-frame' *Pst*I–*Pst*I deletion of 840 bp (lanes 1–3) and the same plasmids with the *Pst*I–*Pst*I deletion and labelled with the *myc*-tag epitope in their C-terminal domain (lanes 4–6) were transfected into COS6 cells. Western blot analysis was performed with anti-CTCF antibodies (**A**), and the membrane stripped and reprobed with the anti-*myc*-tag 9E10 antibodies (**B**). The plasmids with the *Pst*I–*Pst*I deletions, without the *myc*-tag epitope [(A) lanes 1–3, respectively], resulted in a truncated CTCF protein of 97 kDa. The plasmids with the *Pst*I–*Pst*I deletions and the *myc*-tag epitope produced a 110 kDa protein [(A) lanes 4–6, and (B)]. Possible origin of minor multiple bands is discussed in the text. Silent nucleotide substitutions were introduced into the region of the proposed attenuator in pBBc so that the protein sequence remained the same but the poly(A) stretches were disrupted creating pBBc-mut. The COS6 cells were transfected with 10 µg of pBBc and pHHc as controls and pBBc-mutant. Total cell lysates were obtained and Western blot analysis performed as described in Materials and Methods. (**C**) The pBBc and pHHc controls produced 70 and 130 kDa proteins, respectively (lanes 1 and 2) and pBBc-mut produced a 130 kDa protein (lane 3).

plasmids (Fig. 5B, lanes 4–6). Thus the potential terminator of translation is very likely to be located within the *Pst*I–*Pst*I fragment and its removal prevents premature termination.

Point-mutations within the potential attenuator site restore the size of the CTCF-70 encoded in pBBc

To investigate whether the poly(A) sequences within the *Pst*I–*Pst*I region are involved in attenuation, silent point-mutations were introduced into pBBc, creating pBBc-mut, such that they did not change the amino acid sequence, but at the same time disrupted the poly(A) stretches (Fig. 1B). This alteration in the sequence restored the size of the protein expressed *in vivo* to 130 kDa (Fig. 5C, lane 3) suggesting the involvement of these poly(A)-stretches in premature termination.

CTCF-70 may enhance CTCF-130 mediated transactivation in COS6 cells

The ability of one single ORF to produce two alternative products both, *in vivo* and *in vitro*, raises a question of the CTCF-70's functional significance. CTCF-70 can be detected as a minor form in some cell types, but in bursa cells CTCF-70 represents ~25% of the total CTCF proteins (8). As CTCF-70 is correctly located in the nucleus (data not shown), efficiently expressed and corresponds to the CTCF N-terminal domain, one could expect that it may affect the normal function of 130 kDa protein. In fact, biological role of terminal fragments has been previously reported for some transcription factors. In particular, the C-terminal domain of p53 (34) or B-*myc* which represents the N-terminal domain of the *myc* protein (35) can affect transactivation caused by the full length protein. To investigate CTCF-70 ability to interfere with CTCF-130 we employed a standard transactivation assay in COS6 cells as a model system. In these cells co-transfection with pHHc and p90TKLuc [a reporter plasmid containing the dimeric CTCF binding site from the chicken *c-myc* (1,8), inserted in front of a TK Luciferase reporter gene] resulted in transactivation of the luciferase gene (Fig. 6A). Transactivation was dependent on the input of the expressor pHHc. As expected, no transactivation was observed with pBBc in the same assay (Fig. 6B). However, CTCF-70 enhances the ability of CTCF-130 to transactivate the reporter when both expressors are present in COS6 cells (Fig. 6C).

DISCUSSION

The major purpose of the experiments described above was to clarify the relationship between the native CTCF protein and those encoded and produced by the cloned cDNA. Evidence has been provided that: (i) the 130 kDa product is encoded in the 2184 bp ORF which predicts a 82 kDa protein; (ii) this discrepancy is due to an anomalous migration of the protein in the SDS–PAGE gel; (iii) the 70 kDa CTCF protein is a truncated version of CTCF-130 representing its N-terminal domain; and (iv) the 3′-, 5′-UTRs and sequences within the CTCF coding region affect the size of the expressed protein.

The 130 and 82 kDa proteins

The results in this paper demonstrate that the 82 kDa CTCF protein, encoded in the longest ORF of 2.184 kb, migrates as a protein with an apparent molecular weight of 130 kDa in SDS–PAGE, whether the recombinant protein was produced *in vitro* or *in vivo*. This corresponds to the apparent size of the endogenous native CTCF protein. We have also shown that the recombinant proteins are recognised by anti-CTCF antibodies demonstrating that the endogenous and recombinant proteins are identical. The aberrant mobility of the protein can be mainly accounted for by the electrophoretic mobility properties of the N- and C-terminal domains. The anomalous electrophoretic migration of the CTCF C-terminus may be explained by its high negative charge (26% acidic amino acids). Anomalous electrophoretic migration of negatively charged proteins is quite common and has been observed with proteins such as papillomavirus 16 E7 protein (20) and *E.coli ams* protein (36). We also attempted to characterise the CTCF N-terminal domain aberrant migration. However, when pEH (Fig. 1A) was deleted up to the *Bam*HI site and the resulting plasmid pBE expressed in COS6 cells, a 180%

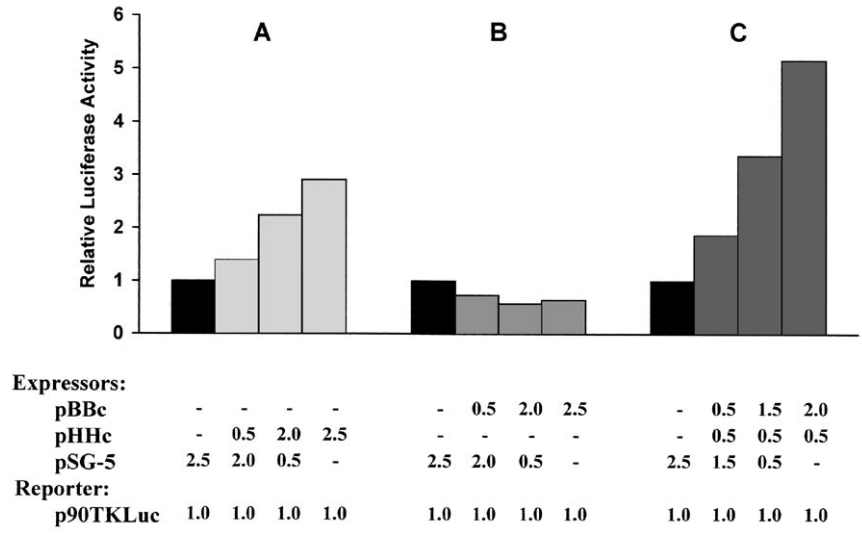

Figure 6. Effect of pBBc on p90-TK Luc transactivation. 1.25×10^5 COS6 cells were plated in 16 mm dish, transfected with Lipofectamine (2.5 µl/dish) and assayed 24 h later. The reporter construct, p90TK -Luc, was co-transfected with expressors (**A**) pHHc, (**B**) pBBc and (**C**) pHHc and pBBc concurrently. The luciferase activities were measured as described in Materials and Methods and are shown in relative units calculated by designating a value of 1 for the background activity observed when p90TK Luc is co-transfected with the 'empty' expression vector pSG5. All other activities were adjusted accordingly. Expression of CTCF-130 and CTCF-70 from pHHc and pBBc were verified by Western blot (data not shown). Each transfection was performed at least three times in duplicate, the average results are shown.

anomaly (21 kDa apparent versus 7.5 kDa theoretical MWs) was still observed. The reason for the anomalous migration of the N-terminal region remains unclear since there is no preponderance of particular types of amino acids.

The 130 and 70 kDa CTCF proteins

We have previously reported (8) that the original CTCF-expressing vectors based on the pBBc construct produced a nuclear protein with apparent molecular weight of 70 kDa *in vivo* and *in vitro* (Fig. 3). The results of this paper demonstrate that: (i) CTCF-70 is a legitimate, not transient, protein product from pBBc since proteins with higher molecular weight from this construct have never been observed, and (ii) its size reduction is due to the loss of the C-terminal part of the CTCF protein. There are at least three possible mechanisms responsible for the appearance of the shorter protein: (i) post-translational processing/splicing of the full length protein; (ii) premature termination of transcription or RNA degradation; and (iii) premature termination of translation of the CTCF mRNA. Mechanism 1 can be ruled out since we were unable to detect a 130 kDa precursor from the pBBc plasmid, after conducting a variety of studies such as pulse–chase experiments, *in vitro* translation, and investigation of a stable cell line producing the 70 kDa protein (8, and unpublished data). Furthermore, proteolysis is considerably unlikely, as no additional bands were detected in extracts from pBBc *myc*-transfectants probed with 9E10 anti-*myc* antibodies (Fig. 4B), and protein splicing is not applicable either, as the entire C-terminal (and, presumably, the Zn finger) domains are missing. Mechanism 2 can be excluded on the basis that with Northern blot analysis, we were unable to detect any additional RNA bands smaller than the 3.7 kb band, which corresponds to the full-length CTCF RNA in various tissues and cell lines including a stable cell line producing the 70 kDa protein (8). Mechanism 3 is quite plausable, as three lines of evidence indicate that CTCF-70 is produced by premature termination of translation: (i) the entire C-terminal part of the CTCF-70 protein is missing as shown by the *myc*-tag approach (Fig. 4); (ii) deletion of the region incorporating the putative attenuator prevents premature termination of translation (Fig. 5A and B); and (iii) point-mutation of the potential attenuator site restores the size of CTCF-70 to 130 kDa (Fig. 5C).

Our investigations demonstrate that only particular combinations of the UTRs with sequences inside the coding region are responsible for the production of CTCF-70. Analysis of the 3′-UTR 792 bp sequence and the coding region revealed the presence of poly(T) and poly(A) stretches, respectively (8). This implies the possibility of an attenuator produced by complementary interaction between the poly(T) and poly(A) stretches. The molecular basis of the production of the CTCF-70 is not known. However, based on data presented in this study and significant information accumulated in the understanding of the general translational termination processes (37–41) we have proposed a model for this mechanism. This model is the basis of ongoing and future investigations of this study and, therefore, seems relevant to this discussion. The involvement of the 5′-UTR and 3′-UTR in the regulation of translation has been convincingly established (29,42,43). Much less is known about elements within coding regions which could contribute to this process. However, it has been shown that sequences within the coding region can affect the mRNA translation efficiency (44,45) and translationally coupled mRNA degradation (46). Our model of the molecular mechanism of premature termination involves both the UTRs and the coding region. In the case of pBBc composed of the ORF and 3′-UTR, the formation of the stem–loop secondary structure between the poly(T) sequences within the 3′-UTR and poly(A) stretches within the putative attenuator, could block the ribosome's passage through the CTCF mRNA. As a result a truncated protein product of 70 kDa is produced. In the case of pEEc, the additional GC-rich 5′-UTR sequences may cause a ribosome 'stalling' at the 5′-end of the CTCFmRNA. This 'stalling', in turn, can lead to destabilisation of the 'downstream' CTCFmRNA stem–loop structure allowing the

ribosome to progress. pHHc, producing CTCF-130, is missing parts of its 5′- and 3′-UTRs. Thus, the pHHc mRNA may not have a developed secondary structure so that translation can be performed without spatial difficulties. This model is supported by the finding that the specific point mutations within the region of the potential attenuator in pBBc (CTCF-70) restored the size of the protein made to 130 kDa. It is likely that the introduced silent nucleotide substitutions altered particular stem–loop formation in the CTCF mRNA so that the ribosome could pass freely.

Functional significance of the 70 kDa protein

The 70 kDa protein may be functionally important and there are indications which make us believe that CTCF-70 does exist under normal physiological conditions. We showed previously that approximately a quarter of total CTCF proteins in bursa was represented by this form (8) and now we demonstrate that CTCF-70 can affect CTCF-130 transactivating capacity in COS6 cells (Fig. 6). Since CTCF-70 does not have the DNA-binding and C-terminal domains, this apparent enhancement may result from the competition of the two forms for a common transcriptional repressor(s). CTCF-70 may have a similar role under normal physiological conditions (e.g. in lymphocytes) providing post-translational tuning of CTCF-130 normal function(s) by competition with interacting protein partners. Three proteins (p32, 80 and 97) have been specifically co-immunuprecipitated with CTCF-130 from COS6 cell extracts, however, one of them, p32, showed interaction with the N-terminal domain (E. Klenova, unpublished observations). A two-hybrid system in yeast is currently being employed to identify and clone other possible CTCF protein partners.

Previous investigations into the existence of the pBBc mRNA under normal physiological conditions included Northern blot analysis from various chicken cell lines and tissues. They revealed two closely migrating bands of ~3.7 and 4.0 kb which are often impossible to resolve on agarose gels (8). It is tempting to speculate that the smaller message corresponds to the mRNA with a part of its 5′-UTR missing (i.e., pBBc). Alternatively, this ‘artificial’ mRNA may serve as an appropriate substrate for the existent molecular apparatus responsible for the similar function in the cell. It seems very intriguing that a message with only 106 bp missing from the 5′-end produces a truncated protein. This may imply a translational control mechanism of protein production carried out by the mRNA itself. In our future studies we will attempt to fully elucidate the physiological significance of this message.

ACKNOWLEDGEMENTS

We thank S. Mittnacht for the anti-GST antibody, C. Sutton for assistance in mass spectrometry, H. Paterson for help with confocal microscopy. We are grateful to K. Davies and I. Craig for helpful discussions. A. Robinson, I. Chernukhin and S. Christodoulou are thanked for critical reading and suggestions on the manuscript. This work was supported by the Human Frontiers Science Program Long-Term Fellowship and The Royal Society Research grant (E.M.K.), Special Research Grant from The Research and Equipment Committee, University of Oxford (R.E.L.), Cancer Research Compaign grants (G.H.G, R.H.N., S.U. and S.C.) and Fogarty International Research Collaboration Award (V.V.L.).

REFERENCES

1 Lobanenkov,V.V., Nicolas,R.H., Adler,V.V., Paterson,H.F., Klenova,E.M., Pololskaya,A.V. and Goodwin,G.H. (1990) *Oncogene* 5, 1743–1753.
2 Filippova,G.N., Fagerlie,S., Klenova,E.M., Myers,C., Dehner,Y., Goodwin,G., Neiman,P.E., Collins,S.J. and Lobanenkov,V.V. (1996) *Mol. Cell. Biol.*, **16**, 2802–2813.
3 Spencer,C.A. and Groudine,M. (1991) *Adv. Cancer Res.* **56**, 1–48.
4 Marcu,K.B., Bossone,S.A. and Patel,A.J. (1992) *Annu. Rev. Biochem.* **61**, 809–860.
5 Kato,G.J. and Dang,C.V. (1992) *FASEB J.* **6**, 3065–3072.
6 Evan,G.I. and Littlewood,T.D. (1993) *Curr. Opin. Gen. Dev.* **3**, 44–49.
7 Lobanenkov,V.V., Adler,V.V., Klenova,E.M., Nicolas,R.H. and Goodwin,G.H. (1990) In Papas,T.S. (ed), *Gene Regulation and AIDS: Transcriptional Activation, Retroviruses and Pathogenesis.* The Portfolio Publishing Co., Texas, pp. 45–66.
8 Klenova,E.M., Nicolas,R.H., Paterson,H.F., Carne,A.F., Heath,C.M., Goodwin,G.H., Neiman,P.E. and Lobanenkov,V.V. (1993) *Mol. Cell. Biol.* **13**, 7612–7624.
9 Grumont,R.J. and Gerondaxis,S. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 4367–4371.
10 Epstein,J.A., Glaser,T., Cai,J., Jepeal,L., Walton,D.S. and Maas,R. (1994) *Genes Dev.* **8**, 2022–2034.
11 Bickmore,W.A., Oghene,K., Little,M.H., Seawright,A., van-Heyningen,V. and Hastie,N.D. (1992) *Science* **257**, 235–237.
12 Molnar,A. and Georgopoulos,K. (1994) *Mol. Cell. Biol.* **14**, 8292–8303.
13 Pulverer,B.J., Kyriakis,J.M., Avruch,J., Nikolakaki,E. and Woodgett,J.R. (1991) *Nature* **353**, 670–674.
14 Reason,A.J., Morris,H.R., Panico,M., Triesman,R.H., Haltiwanger,R.S., Hart,G.W., Kelly,W.G. and Dell,A. (1992) *J. Biol. Chem.* **267**, 16911–16921.
15 Eitan,S. and Schwartz,M. (1993) *Science* **261**, 106–108.
16 Roth,M.J., Brown,D.R. and Hurwitz,J. (1984) *J. Biol. Chem.* **259**, 10556–10568.
17 Challberg,M.D., Desiderio,S.V. and Kelly,T.J. (1990) *Proc. Natl. Acad. Sci. USA* **77**, 5105–5109.
18 Lee,Y.F., Nomoto,A., Detjen,B.M. and Wimmer,E. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 59–63.
19 Fontoura,B.M.A., Sorokina,E.A., David,E. and Carroll,R.B. (1992) *Mol. Cell. Biol.* **12**, 5145–5151.
20 Armstrong,D.J. and Roman,A. (1993) *Biochem. Biophys. Res. Commun.* **192**, 1380–1387.
21 Gopal,V., Ma,H-W., Kumaran,M.K. and Chafferji,D. (1994) *J. Mol. Biol.* **242**, 9–22.
22 Pham,D.Q.D. and Sivasubramanian,N. (1992) *Gene* **122**, 345–348.
23 Query,C.C., Bentley,R.C. and Keene,J.D. (1989) *Cell* **57**, 89–101.
24 Blank,V., Kourilsky,P. and Israel,A. (1991) *EMBO J.* **10**, 4159–4167.
25 Whiteside,S.T., King,P. and Goodbourn,S. (1994) *J. Biol. Chem.* **269**, 27059–27065.
26 Cooper,A.A. and Stevens,T.H. (1993) *BioEssays* **15**, 667–674.
27 Yanofsky,C.J. (1988) *Biol. Chem.* **263**, 609–612.
28 Spencer,C.A. and Groudine,M. (1990) *Oncogene* **5**, 777–785.
29 Sonenberg,N. (1994) *Curr. Opin. Gen. Dev.* **4**, 310–315.
30 Roth,M.B., Zahler,A.M. and StolkJ.A. (1991) *J. Cell. Biol.* **115**, 587–596.
31 Nordeen,S.K. (1988) *Biotechniques* **6**, 454–457.
32 Smith,D.B. and Johnson,K.S. (1988) *Gene* **67**, 31–40.
33 Evan,G.I., Lewis,G.K., Ramsay,G. and Bishop,J.M. (1985) *Mol. Cell. Biol.* **5**, 3610–3616.
34 Shaulian,E., Haviv,I., Shaul,Y. and Oren,M. (1995) *Oncogene* **10**, 671–680.
35 Resar,L.M.S., Dolde,C., Barrett,J.F. and Dang,C.V. (1993) *Mol. Cell. Biol.* **13**, 1130–1136.
36 Casarègola,S., Jacq,A., Laoudj,D., McGurk,G., Margarson,S., Tempête,M., Norris,V. and Holland,I.B. (1992) *J. Mol. Biol.* **228**, 30–40.
37 Kozak,M. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 2850–2854.
38 Tuite,M.F. and Stansfield,I. (1994) *Mol. Biol. Rep.*, **19**, 171–181.
39 Hershey,J.W.B. (1991) *Annu. Rev. Biochem.*, **60**, 717–755.
40 Stansfield,I., Jones,K.M. and Tuite,M. (1995) *Trends Biol. Sci.* **20**, 489–491.
41 Nakamura,Y., Ito,K. and Isaksson,L.A. (1996) *Cell* **87**, 147–150.
42 Jackson,R.J. and Standart,N. (1990) *Cell* **62**, 15–24.
43 Standart,N. and Jackson,R.J. (1994) *Biochimie*, **76**, 867–879.
44 Milland,J., Christiansen,D., Thorley,B.R., Mckenzie,I.F.C. and Loveland,B.E. (1996) *Eur. J. Biochem.* **238**, 221–230.
45 Ruiz-Linares,A., Bouloy,M., Girard,M. and Cahour,A. (1989) *Nucleic Acids Res.* **17**, 2463–2476.
46 Schiavi,S.C., Wellington,C.L., Shyu,A-B., Chen,C.-Y., Greenberg,M.E. and Belasco,J.G. (1994) *J. Biol. Chem.* **269**, 3441–3448.