The EMBO Journal vol.4 no.6 pp.1609–1614, 1985

Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus* oocytes

J.Miller, A.D.McLachlan and A.Klug

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

Communicated by A.Klug

The 7S particle of *Xenopus laevis* oocytes contains 5S RNA and a 40-K protein which is required for 5S RNA transcription *in vitro*. Proteolytic digestion of the protein in the particle yields periodic intermediates spaced at 3-K intervals and a limit digest containing 3-K fragments. The native particle is shown to contain 7–11 zinc atoms. These data suggest that the protein contains repetitive zinc-binding domains. Analysis of the amino acid sequence reveals nine tandem similar units, each consisting of approximately 30 residues and containing two invariant pairs of cysteines and histidines, the most common ligands for zinc. The linear arrangement of these repeated, independently folding domains, each centred on a zinc ion, comprises the major part of the protein. Such a structure explains how this small protein can bind to the long internal control region of the 5S RNA gene, and stay bound during the passage of an RNA polymerase molecule.

Key words: Xenopus laevis/transcription factor IIIA/zinc-binding domains/7S particle

Introduction

The 5S RNA genes of *Xenopus laevis*, which are transcribed by RNA polymerase III, have been the subject of intensive study in the last decade by the research groups of D.D.Brown and R.G. Roeder. The two types, oocyte and somatic, provide a system for studying the differential regulation of gene expression as well as transcription mechanisms (Brown, 1984). In the course of these studies it has been discovered that the correct initiation of transcription requires the binding of a 40-K protein factor, variously called factor A or transcription factor IIIA (TFIIIA), which has been purified from oocyte extracts (Engelke *et al.*, 1980). By deletion mapping it was found that this factor interacts with a region ~50 nucleotides long within the gene, called the internal control region (Bogenhagen *et al.*, 1980). This initiation complex is stabilised by the sequential binding of two further protein factors, called B and C (Segall *et al.*, 1980; Bieker *et al.*, 1985).

Immature oocytes store 20 000 5S RNA molecules in the form of 7S ribonucleoprotein particles (Picard and Wegnez, 1979), each containing a single protein which has been shown to be identical with transcription factor IIIA (Pelham and Brown, 1980; Honda and Roeder, 1980). TFIIIA therefore binds both 5S RNA and its cognate DNA and it was therefore suggested that it may mediate autoregulation of 5S gene transcription. Whether this autoregulation occurs *in vivo* or not, the dual interaction provides an interesting structural problem which can be approached because of the presence of large quantities of the protein TFIIIA in *Xenopus* oocytes. In this paper we report some results of our preliminary studies on TFIIIA which reveal a remarkable repeating structure within the protein.

Results

Zinc content of 7S particle

Because published methods for 7S purification involve ion exchange and high ionic strength buffers, which in our hands dissociates at least 30% of particle samples, we developed a new method for particle purification which causes no detectable dissociation. We found, in agreement with other workers (Denis and le Maire, 1983), that the 7S complex dissociated at salt concentrations >0.2 M, and that the isolated protein precipitated readily. We also observed that gel filtration of the complex in the presence of 0.1 mM dithiothreitol (DTT) invariably resulted in separate elution of protein and 5S RNA. This dissociation was also seen when the particle incubated with 0.1 mM DTT was electrophoresed on agarose gels, suggesting that the 20 cysteine groups per molecule (Picard and Wegnez, 1979) were somehow involved in particle stability. However, when we found that 25 mM $NaBH_4$ did not disrupt the complex, we suspected, by analogy with the results of Lewis and Laemmli (1982) on metaphase chromosomes, that metal binding might be involved. When the particle was incubated with 1,10-phenanthroline, DTT, EDTA or a number of other chelating agents and run on agarose gels, dissociation was observed which could only be prevented by prior addition of Zn^{2+}, and not by Mg^{2+}, Ca^{2+}, Mn^{2+}, Ni^{2+}, Fe^{2+}, Ca^{2+} or Co^{2+}. Cu^{2+} and Cd^{2+} induced dissociation by themselves, apparently by displacing bound Zn^{2+}.

Analysis of a 30–50% pure preparation of particle by atomic absorption spectroscopy revealed insignificant concentrations of Cd, Cu, Ni, Co or Fe, but a significant concentration of Zn, at a ratio of at least 5 mol Zn per mol. particle.

While these experiments were in progress, Hanas *et al.* (1983) reported the presence of Zn in the 7S RNP particle, and the requirement for Zn for the association of TFIIIA with the internal control region of the 5S gene. They found a Zn/particle ratio of about two in the presence of 5 mM EDTA and three in its absence, although the latter fell to two in samples purified by gel filtration.

We have now repeated the analysis with pure and undissociated particle preparations. To ensure that no contamination of the preparation could occur, all glass and plastic ware was washed in several changes of 10 mM EDTA for a time several times longer than the duration of the entire preparative procedure. The buffer was concentrated and submitted with particle sample for atomic absorption spectrophotometry. 7S particle at 65 μM contained Zn at 460 μM, giving a ratio of 7.0 ± 0.5 mol Zn per mol particle. The original buffer used was at most 60 nM in Zn, and particle was exposed to at most 1.5 litres of buffer during the purification. Since the yield of particle was 2 ml solution at 65 μM, containing 920 nmol of Zn, at most 90/920 = 10% could have been adsorbed at any time following homogenization of frog ovaries.

The buffers used by Hanas *et al.* (1983) contained 0.5 mM or 1 mM DTT, which has a large binding constant for Zn of ~10^{10} (Cornell and Corviro, 1972), and so their value for the Zn content may be an underestimate. Our buffer contained

20 mM MES, a weakly chelating buffer, which nevertheless, because of the large volumes used in the gel filtration and dialysis steps in the preparative procedure, might have reduced the Zn content of the particles, suggesting that the value we have obtained of seven Zn/particle may still be an underestimate. An experiment in which the particle was prepared in the presence of 10 μM Zn and then separated from unbound Zn on a gel filtration column gave a value of 11 – 12 mol Zn/mol particle. The sequence analysis described below suggests that there may be at least 9 mol Zn/mol particle.

The 7S particle is not unique in its requirement for Zn. The 42S particle of *X. laevis*, which Denis and le Maire (1983) have shown to contain, among other components, a 5S RNA and a 5S RNA-binding protein distinct from TFIIIA, also requires Zn for stability by agarose gel assay, has a very large molar Zn content, and contains no significant amounts of other metals (Miller, unpublished results).

Small domains in TFIIIA protein

Smith *et al.* (1984) have shown that, on treatment with proteolytic enzymes, the 40-K intact protein gives rise to a 30-K breakdown product, which is then converted to a 20-K product. These proteolytic fragments remain bound to the 5S RNA but Smith *et al.* (1984) have purified them and studied their interactions with the 5S gene by DNase I footprinting. From these experiments, Smith *et al.* (1984) concluded that TFIIIA consists of three structural domains which bind to different parts of the internal control region of the 5S gene: a 20-K protein which binds to the 3′ end of the coding strand, an adjacent 10-K domain which extends the binding to the 5′ end of the control region, and a third domain of 10-K which does not bind directly to the DNA but enables the intact protein to enhance transcription, presumably through interaction with RNA polymerase.

We have also carried out proteolysis of the 7S particle using trypsin, chymotrypsin, elastase, and also papain (D.Rhodes, personal communication) to determine whether smaller fragments of the protein might bind the 5S RNA and the 5S RNA gene. Our measurements of the tryptic breakdown products suggest that Smith's 40-K, 30-K and 20-K may be closer to 39-K, 33-K and 23-K (data not shown).

We have also observed that the 23-K fragment may be reduced to a 17-K fragment which remains bound to the 5S RNA. Further, the 17-K fragment, after prolonged proteolysis, can be reduced to a limit digest consisting of a mixture of 6-K, 4-K and 3-K fragments (Figure 1), which ultimately themselves disappear. Chymotrypsin produces higher multiples of these fragments as well. Doublets are seen at ~11 and 9.5-K, 7.5 and 6-K and 4-K and 3-K, a spacing between them of 3 – 3.5-K (Figure 2c and d). Early time points in tryptic and chymotryptic digestion reveal metastable bands between the production of the major fragments described above (data not shown), which are also approximately periodically spaced.

The finding of periodic intermediates in the course of proteolytic digestion, and the persistence of small fragments even after prolonged digestion, suggests a periodic arrangement of small, compact protein domains of size ~3-K. If such repetitive domains existed they might account for the large number of cysteine and histidine residues and the multiple Zn binding.

We therefore investigated the newly published amino acid sequence (Ginsberg *et al.*, 1984) to see if any structural periodicity was manifested in the sequence.

Analysis of sequence of TFIIIA

Sequence repeats. The amino acid sequence of TFIIIA, which has been deduced from a cDNA clone (Ginsberg *et al.*, 1984), contains an unusually large number of Cys and His residues. At first sight these residues appeared to us to form roughly periodic groupings. We therefore made a systematic search for repeats in both the amino acid sequence and the cDNA, using the diagonal comparison matrix method and the damped Needleman and Wunsch method (see Materials and methods).

The protein comparison matrices (with short window lengths of 11, 22 and 30 residues) showed an exceptionally strong and regular pattern of 30-residue repeats in the sequence, with four repetitions evident in the first half of the molecule (residues 13 – 156) and two more clear repeats in the second half (residues 223 – 276). Further analysis of the protein by the damped Needleman and Wunsch method showed that the repeats were even more extensive, being partly obscured by a few gaps in the middle of the sequence. The final best alignment (Figure 3) shows that

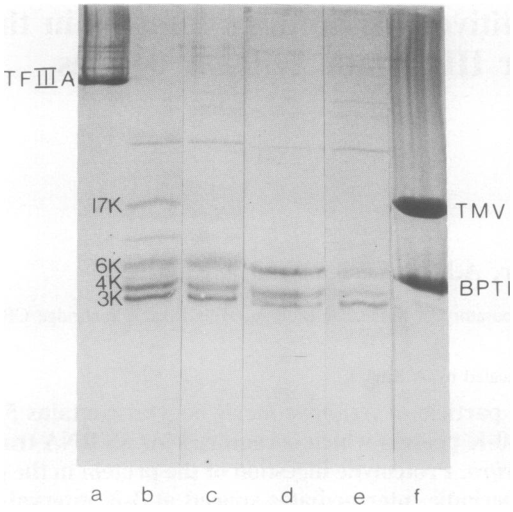

Fig. 1. Trypsin digestions. Particle samples (0.2 mg/ml) in 20 mM MES buffer, pH 6.0, 50 mM KCl were dialyzed against 50 mM Tris-Cl, pH 8.1, 50 mM KCl. 20 μl aliquots were digested with trypsin (20 μg/ml) for varying times and stopped with 2 mM benzamidine. Times were **(a)** 0; **(b)** 17 h; **(c)** 24 h; **(d)** 39 h; **(e)** 64 h. Electrophoresis was as described in Materials and methods.

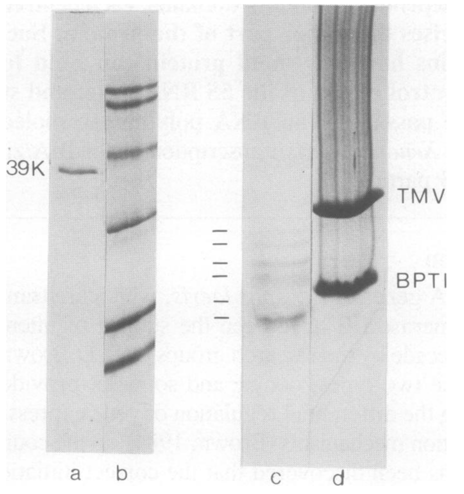

Fig. 2. (a) TFIIIA used in these experiments; **(b)** markers: 12.3 K, 17.2 K, 25.7 K, 45 K, 66.3 K, 78 K; **(c)** 17 h chymotrypsin (20 μg/ml) digest of sample prepared as described in legend to Figure 1. Doublet bands are marked (see Results). Acrylamide concentrations for **(a)** and **(b)** were 15%.

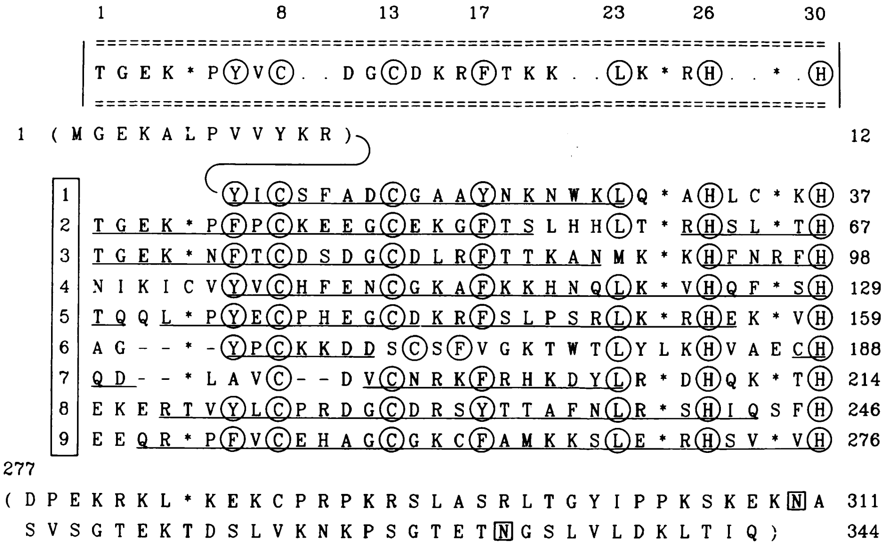
1 8 13 17 23 26 30
T G E K * P Y V C . . D G C D K R F T K K . . L K * R H . . * . H
1 (M G E K A L P V V Y K R) 12
1 Y I C S F A D C G A A Y N K N W K L Q * A H L C * K H 37
2 T G E K * P F P C K E E G C E K G F T S L H H L T * R H S L * T H 67
3 T G E K * N F T C D S D G C D L R F T T K A N M K * K H F N R F H 98
4 N I K I C V Y V C H F E N C G K A F K K H N Q L K * V H Q F * S H 129
5 T Q Q L * P Y E C P H E G C D K R F S L P S R L K * R H E K * V H 159
6 A G - - * - Y P C K K D D S C S F V G K T W T L Y L K H V A E C H 188
7 Q D - - * L A V C - - D V C N R K F R H K D Y L R * D H Q K * T H 214
8 E K E R T V Y L C P R D G C D R S Y T T A F N L R * S H I Q S F H 246
9 E E Q R * P F V C E H A G C G K C F A M K K S L E * R H S V * V H 276
277
(D P E K R K L * K E K C P R P K R S L A S R L T G Y I P P K S K E K N A 311
S V S G T E K T D S L V K N K P S G T E T N G S L V L D K L T I Q) 344

Fig. 3. Amino acid sequence of transcription factor IIIA from *X. laevis* oocytes, aligned to show the repeated units. The sequence is in one-letter code (Dayhoff, 1978). The molecule contains: an amino end region (residues 1 – 12); a lysine-rich zone (277 – 309) near the carboxyl end; a short tail region which may bind carbohydrate at Asn 310 and Asn 333, which are indicated by squares. The repeat units are numbered 1 – 9 on the left side of the diagram. The boxed-in consensus sequence at the top shows the characteristic features of a typical repeat unit, numbered as for a length of 30 residues. The end-point of each unit has been chosen arbitrarily after His-30. The best-conserved residues are ringed. Cys-8, Cys-13, His-26 and His-30 are believed to bind to a Zn^{2+} ion. Tyr-6, Phe-17 and Leu-23 are hydrophobic residues which may form an inner core for the proposed multiple-domain structure. Asterisks (*) mark positions where an insertion sometimes occurs in the normal pattern, and dots (.) mark variable positions in the sequence. In the main body of the repeats a dash (-) indicates an alignment gap. The underlined regions are those which show clear evidence of a relationship with at least one other unit (see Table I). Note that units 6 and 7 have diverged considerably from the usual pattern, and that residues 277-282 may form a fragmentary extension of repeat unit 9.

Table I. Overall similarities between the nine repeat units

Residues	Unit	Similarity values								
		5	9	2	8	3	4	1	7	6
130 – 159	5	*	23	20	15	18	14	6	8	2
247 – 276	9	23	*	23	16	18	12	10	1	0
38 – 67	2	20	23	*	11	17	16	6	1	8
215 – 246	8	15	16	11	*	17	13	11	5	5
68 – 98	3	18	18	17	17	*	6	6	6	2
99 – 129	4	14	12	16	13	6	*	16	7	0
13 – 37	1	6	10	6	11	6	16	*	0	0
189 – 214	7	8	1	1	5	6	7	0	*	1
160 – 188	6	2	0	8	5	2	0	0	1	*
Column totals		106	103	102	93	90	84	55	28	18

The value given to each pair is the number of overlapping 11-residue windows, in the correctly aligned comparison of the two units, for which the score exceeds the 0.1% matching probability level. Since each unit is ~30 residues long, the maximum possible value is 30. The column totals give a measure of how close each unit is to the global consensus of them all. The underlinings below the sequence in Figure 1 show the parts of each unit that lie in the centre of a window, where the comparison score, with at least one other unit, reaches the 0.1% level.

residues 13 – 276 form a continuous run of a repeated motif, of nine similar units, each of ~30 residues. Since the sequence repeats itself cyclically, the position of the boundary of the motif is uncertain, but the choice of an end-point shortly after His-30 of the consensus allows the largest number of complete units to be assigned. With this choice, most of the insertions and deletions occur near the ends of the units rather than in the central loop (consensus position numbers 14 – 25).

Our results suggest strongly that residues 13 – 276 of TFIIIA have evolved from a primitive ancestral unit of 30 amino acids by gene duplication or gene conversion (Hood *et al.*, 1975). It is not yet possible to make a proper analysis of the evolutionary relationships, because the end points of the units are not known; if the TFIIIA gene possesses intervening sequences these could be revealing. Table I gives a rough estimate of the degree of relatedness of each pair of units, derived from the numbers of high-scoring windows in the comparison matrix. The totals of the columns in the table give a measure of how closely each unit resembles the global consensus of the nine repeats.

It can be seen that units 5, 9 and 2, in that order, are clearly the most typical members of the family, and are very like one another. Common features include Pro-5, Gly-12, Lys-15 and Arg-25 (here we refer to the numbering of the consensus sequence in Figure 3). A second group comprises units 8, 3 and 4. These contain some single-residue insertions at positions 4A or 28A. Units 8 and 3 form a recognisable pair as shown by the identities Asp-11, Gly-12, Asp-15, Thr-18, Thr-19, Asn-22 and Phe-29. Lastly, units 1, 7 and 6 have diverged considerably from the normal pattern: in particular, units 6 and 7 are shortened at their amino ends with irregular (Cys-8)-(Cys-13) loops. We note that units 1 and 4 have Phe-10, Gly-14, Ala-16 and Lys-19 in common, while units 1 and 6 both have Trp-21.

Thus although all nine units belong to the same family they have diverged in detail and may have taken on specific individual DNA-binding functions as the 5S gene control system evolved. It would also not be surprising if the same 30-residue units were later found to occur in varying numbers in other related gene control proteins.

The evolutionary advantages of a repeating design are probably much the same as those in many other linear proteins, such as ovomucoid inhibitor (Laskowski *et al.*, 1980). Probably a single functional unit that binds to a half-turn of DNA was once evolved, and then became used in much more subtle and specialised ways when a large number of similar units were joined in series.

Characteristic structural features of the small domains

The well-marked repeat pattern in TFIIIA suggests strongly that each of the sequence repeats corresponds to a series of small structural elements, or domains, of ~30 residues, arranged linearly. In large globular proteins with repeated sequences the units most often form compact dimeric or tetrameric pseudo-symmetrical structures (Rossman and Argos, 1981; McLachlan, 1980). But there are also well-known examples of proteins with compact independently active structural units strung together in line, which can be separated by enzyme cleavage: the 'kringles' in plasminogen (Sottrup-Jensen *et al.*, 1978) and the Kazal-type protease domains in the ovoinhibitor of Japanese quail (Laskowski *et al.*, 1980; Bolognesi *et al.*, 1982). In these molecules the structural units have lengths of ~80 and 63 residues. The TFIIIA unit is exceptional because of its small size and unique arrangement of Cys . . . Cys . . . His . . . His residues.

There is strong evidence, both that the repeat unit is a self-sufficient folded domain, and that it is stabilised by zinc ions. The first conclusion is suggested by the appearance of multiple small fragments on proteolysis, with quantized mol. wts. which are often multiples of 3 kd (see above). The second is supported by evidence that one TFIIIA molecule binds many zinc ions (7–11, see above) and is inactive in the absence of the metal (Hanas *et al.*, 1983). We therefore believe that most, if not all, of the nine units bind zinc.

Zinc is normally tetrahedrally coordinated in inorganic and metallo-organic compounds. The amino acids Cys and His are its most common ligands in enzymes (see Fersht, 1977), such as carbonic anhydrase, liver alcohol dehydrogenase, carboxypeptidase A, Cu-Zn superoxide dismutase and thermolysin. Therefore we may picture each small domain as a compact unit formed round a central zinc ion to confer stability. Each zinc would be coordinated tetrahedrally to the two invariant pairs of Cys and His residues in each unit. The ends of the domain will then be pulled together round the zinc. In Figure 4 we have drawn one possible arrangement of the zinc ligands which fits this type of scheme. It is important to remember that the Cys . . . Cys and His . . . His loops probably cross over at an angle to form a tetrahedral box, and that the zinc may have more than four ligands. For example, a Cys side chain might be shared between zincs in two adjacent units.

Figure 5 shows how each domain contains three invariant hydrophobic groups (normally Tyr-6, Phe-17 and Leu-23) and how the acidic side chains are nearly all in or near the (Cys-8) . . . (Cys-13) loop.

The extended protein loop between residues 14 and 25 might form a DNA-binding region. The three-dimensional structures of gene repressor proteins (Ohlendorf and Matthews, 1983) show that the side chains of Lys, His, Asn, Gln and Thr often interact with the phosphate backbone of DNA, while Arg can form hydrogen bonds to the base pairs. Amino acids of this type are concentrated in the region 14–25 of TFIIIA (Figure 4).

A structural parallel for our proposed 30-residue metal-centered domain might be the 40-residue calcium-binding unit of the calmodulin family (Moews and Kretsinger, 1975) or the 26-residue ion-sulphur half-domain of bacterial ferredoxin (Adman *et al.*, 1973). However, neither of these structures is an independent folding unit, but forms part of a linked pair of fragments. The metallothionein family of proteins (Boulanger *et al.*, 1983), which contain repeated Cys residues, are also not a good analogy for TFIIIA, but a rather peculiar group. The reason is that they bind metal ions in clusters, bridged by Cys residues, so that the ions are not separately packaged in independent protein cages, but instead built into a 'metal-sulphide' pseudocrystal bordered by the polypeptide chain.

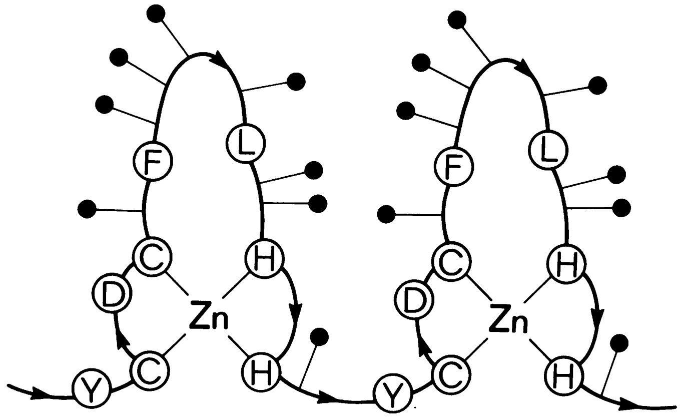

Fig. 4. Folding scheme for a linear arrangement of repeated domains, each centred on a tetrahedral arrangement of zinc ligands. Ringed residues are the conserved amino acids which include the Cys and His zinc ligands, the negatively charged Asp-11, and the three hydrophobic groups that may form a structural core. Black circles mark the most probable DNA-binding side chains (see Figure 5). In the scheme drawn here the metal ion draws the ends of each unit together, leaving the central residues 14–25 to form a potential DNA-binding loop or 'finger'. An alternative but much less likely position for the zinc is between the His residues of one unit and the Cys residues of an adjacent one.

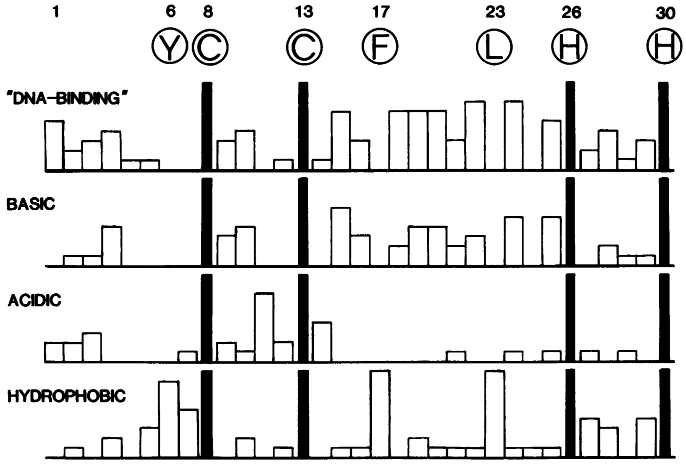

Fig. 5. Histograms for the average distribution of amino acids along the length of each 30-residue repeat unit. The height of each bar, in the range 0–9, gives the number of times that class of amino acid occurs at each position in the nine units. Positions are numbered as in Figure 3, with Cys-8, Cys-13, His-26 and His-30 treated as special positions (zinc ligands). Amino acids have been classed as follows. (a) DNA-binding: Lys, Arg, His, Asn, Gln, Thr (Ohlendorf and Matthews, 1983). (b) Basic: Arg, Lys, His. (c) Acidic: Asp, Glu. (d) Large hydrophobic: Leu, Ile, Val, Phe, Tyr, Trp. The strongest potential DNA-binding sites, with five or more DNA-binding amino acids in the nine units, are also marked in Figure 4.

TFIIIA thus appears to have an exceptionally compact molecular architecture for a self-sufficient structure; however, it may be that strong interactions between adjacent units are important for maintaining the structure. We have examined the two-helix DNA-binding motif of the Cro and Lac repressors, but it does not seem possible to accommodate the TFIIIA consensus sequence into this model: the loop 14–25 may instead fold into a long twisted ribbon of β-sheet which wraps in some way round the zinc-binding pocket, using the invariant hydrophobic groups as nuclei. The characteristic Cys-Cys-His-His consensus motif

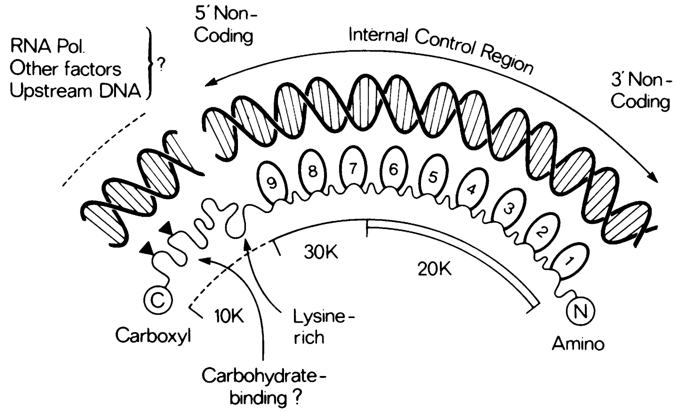

Fig. 6. An interpretation of the structural features of the protein TFIIIA and its interactions with DNA. The DNA is drawn curved, as if resting on a long beaded surface of the protein. The internal control region of the 5S RNA gene (bases 40 – 100) is drawn as six turns of DNA, with the 5′ end of the non-coding strand at the left. A separate piece of upstream DNA may be required to promote transcription. The protein sequence runs from right to left: the evidence for this orientation is given in the text. The amino end is followed by nine repeat units (residues 1 – 276) in contact with the control region. These units together form the 30-K proteolytic fragment of Smith *et al.* (1984). The first six repeat units (residues 13 – 188) probably constitute the 20-K proteolytic fragment: irregularities in the repeat pattern of units 6 and 7 may correspond to a susceptible cleavage region between the two fragments. The repeats are thought to be extended DNA-binding 'fingers' linked by flexible joints, each having a zinc ion centre. Residues 277 – 344 include a lysine-rich region, followed by two potential carbohydrate-binding sequences near the carboxyl end (marked with filled triangles), and together these parts form a separate domain, the 10-K proteolytic fragment. The relationship between the grooves of the DNA and the positions of the protein units is unknown and the drawing must not be taken literally, but each unit binds to about one half-turn of DNA. The tip of a unit could fit into a groove.

found in TFIIIA appears to be unique to this protein. We have searched for it without success in a large number of zinc enzymes, including those mentioned above, as well as in the metallothioneins. There is, therefore, no evidence yet that this motif, or indeed any other repetitive pattern, is a typical feature of zinc-binding proteins as a class.

Relation of large domains to the protein sequence. The nine small domains containing repeated pairs of cysteines and histidines comprise residues 1 – 276 or 1 – 280 (depending where the boundary is taken). The remaining 70 residues at the carboxyl end have no homology with the repeating units (Figure 3). The sequence suggests that this terminal region might be composed of two parts. The first half of ~30 residues is very rich in lysines and arginines, resembling in this respect a histone, although no significant sequence homology emerges. The second half lacks this enrichment and, as has been pointed out to us by Dr H-C. Thøgersen, contains two potential sites for carbohydrate addition at aspargines 310 and 333. [The characteristic sequence is N X S (Marshall, 1972).]

The marked difference in character of the last 70 residues suggest that this region might be identified with the 10-K 'domain' revealed at one end of the protein by the proteolysis studies of Smith *et al.* (1984). This 10-K domain is required for efficient RNA transcription, but, unlike the remaining 30-K fragment (and its smaller 20-K subfragment), not for binding to the internal control region of the 5S RNA gene. The amino and carboxyl orientation of the protein relative to the gene would then be as shown in Figure 6.

This orientation is consistent with the cyanogen bromide cleavage map of the protein shown by Smith *et al.* (1984, Figure 2). As pointed out by them, the small fragment CB2 stains quite differently with silver from the two other fragments. We therefore identify the fragments CB3, CB1 and CB2 with the residues 1 – 90, 91 – 266 and 267 – 344, respectively, since the latter is of different character from the first two (particularly in its high content of lysine). The lower mobility of the shorter peptide CB3 (78 residues) relative to CB3 (90 residues) could be accounted for if some of its residues were modified, as suggested above. The question of modifications is being investigated. The same amino-carboxyl orientation (Figure 6) has been found by the Carnegie group (D.D.Brown, personal communication).

Discussion

The experimental and theoretical studies described above have led to a picture in which the transcription factor IIIA consists mostly of a linear arrangement of nine repeated domains, each centred on a tetrahedral arrangement of zinc ligands. The repeats are thought to be extended DNA-binding fingers, linked by (flexible?) joints. This model is consistent with the highly asymmetric shape of the molecule indicated by its physico-chemical properties (Bieker and Roeder, 1984). A structure of this kind would explain how a relatively small protein of 40 K can bind to a long stretch of double-helical DNA: if each domain interacted with about half a DNA period, then, allowing for end effects, the nine domains could cover ~50 – 60 nucleotides, the length of the internal control region (Sakonju *et al.*, 1980). The DNA or RNA binding strength of each domain could be modulated, and specific recognition of short nucleic acid stretches established, by variations in the sequence at the finger tips.

One advantage of a many-fingered design for the transcription factor is that it binds to an internal control region of the gene in a system where a stable transcriptional complex is formed (Bogenhagen *et al.*, 1982; Gottesfeld and Bloomer, 1982; Lassar *et al.*, 1983). This complex can sustain many rounds of transcription during which the factor presumably remains bound to the gene. As the polymerase passes through the gene, the many-fingered protein could release those fingers bound ahead of the processing polymerase, but stay bound by its remaining fingers, whether to the intact DNA double helix or to the non-coding strand.

We have already mentioned above that the transcription factor appears to be a highly evolved version of a small molecule of the size of one of the contemporary 3-K domains stabilised by a metal ion. The primitive molecule could have simply assisted an early form of transcription. Evolution to the elaborate transcription apparatus found today could have taken place by gene duplication, with different repeats taking up extra functions. It is noteworthy, as remarked by Hanas *et al.* (1983) that RNA polymerase III contains zinc, and it could be that the initial activity of primitive TFIIIA promoted transcription in the absence of a polymerase molecule, which presumably only evolved later for greater efficiency.

Ginsberg *et al.* (1984) have noted a homology between a region of 5S DNA (or RNA) and the coding sequence of the TFIIIA gene, and have suggested that TFIIIA could interact with its own gene or the derived RNA to autoregulate expression at the transcriptional or translational levels. The evolutionary origin of this contemporary property may be as follows. RNA is widely believed to have preceded DNA, so that a small primitive RNA could have coded for a small protein (the precursor of the 3-K repeating unit) which bound back to the RNA and so stimulated its own production.

It may be that the deletions in units 6 and 7 of the contemporary protein (Figure 3) are necessary to enhance its flexibility and enable it to accommodate to the secondary or tertiary structure of the 5S RNA, as well as to the DNA.

The 10-K domain of Smith *et al.* (1984) does not bind to the internal control region of the gene and we have identified it with the 70 residues at the C-terminal end of the protein. Smith *et al.* (1984) have suggested that this 'transcription' domain may interact with other factors required to form a competent transcription complex or directly with RNA polymerase III. To this we would only add that, in view of its high lysine and arginine content, it might also be involved in binding to upstream elements of the DNA outside the coding region of the gene.

Materials and methods

7S particle purification

7S RNP particles were prepared from the ovaries of immature *X. laevis* (South African Snake Farm) by a method to be described elsewhere (Miller, in preparation). The method does not subject the particle to strong chelators, such as DTT, or to salt concentrations above 0.07 M. Protein is at least 95% pure by Coomassie or silver stain, and >90% by amino acid analysis (data not shown). The particle is no more than 5% dissociated as measured by agarose gel electrophoresis (see below) and usually contains no detectable free 5S RNA.

For quantitation of metal content, glassware, tubing and any item to which buffer was exposed was treated for at least 48 h in 30% nitric acid, washed extensively with glass-distilled water and then treated for at least 72 h in 10 mM Na_2 EDTA (pH 4), the optimum pH for Zn^{2+} complexation by EDTA (West, 1969). Any items not resistant to nitric acid, such as column matrices or dialysis tubing, were rinsed in several changes of 10 mM Na_2 EDTA (pH 4) for at least 72 h. All materials were washed with glass-distilled water before use. Buffers were passed through a Chelex 100 ion-exchange column (BioRad) directly before use to remove divalent cations. Protein concentration was measured by amino acid analysis.

Particle dissociation assay

7S particle preparations, usually at concentrations of 100 μg/ml, were loaded with an equal volume of 50% glycerol, 50 mM Tris-Cl, pH 7.5, 0.1% xylene cyanol onto 0.7% agarose gels in 50 mM Tris-Cl, 90 mM boric acid pH 8.5 and run for ~1 h at 8 V/cm. Gels were stained with 1 mg/ml ethidium bromide and visualised under u.v. Experiments have confirmed that the fraction of RNA in particle preparations co-migrating with free 5S RNA correlates qualitatively with the fraction of 5S RNA eluting in a separate peak from 7S particle in gel filtration (Miller, unpublished results), suggesting that it may be taken as a measure of free and bound 5S RNA in solution.

Polyacrylamide gel electrophoresis

Electrophoresis was as described by Laemmli (1970) with the following modifications. Stacking gel was 3% acrylamide, 0.15% bis-acrylamide, 0.125 M Tris-PO_4, pH 6.8, 0.1% SDS, 0.01% TEMED, 0.1% ammonium persulphate. Running gel was 22.5% acrylamide, 0.73% bis, 0.75 M Tris-Cl pH 8.8, 0.1% SDS, 0.01% TEMED, 0.1% ammonium persulphate. Samples were diluted with an equal volume of 100 mM $NaPO_4$ pH 6.8, 2% SDS, 100 mM DTT and 50% glycerol and boiled for 3 min directly before loading. Approximate mol. wts. were found from a logarithmic plot of mobility against the mol. wts. of two markers, tobacco mosaic virus protein (TMV) and bovine pancreatic trypsin inhibitor (BPTI). Gels were stained with 0.5% PAGE blue 83 (BDH Biochemicals).

Proteolysis

Proteolysis was always conducted at particle concentrations of 200 – 500 mg/ml in 50 mM Tris-Cl pH 8.1, 50 mM KCl, 0.5 mM $MgCl_2$, with 20 μg/ml trypsin, room temperature. Reactions were stopped either with 2 mM benzamidine, or by boiling for 2 min in loading buffer.

Sequence repeats – methods of analysis

We used two well-established methods. In the first, the diagonal comparison matrix (McLachlan, 1971, 1983), the sequence is divided into all its possible overlapping segments, or windows, of a given fixed length, without insertions or deletions, and every pair of segments is compared independently. The score for comparing each pair of amino acids in the window is derived from Dayhoff's mutation likelihood tables (Dayhoff, 1978) and is assessed against the exact calculated probability distribution of the window scores for random sequences with the same average composition as the whole protein (McLachlan, 1983). The cDNA sequences were scored by counting identical bases. In the second method, the damped Needleman-Wunsch alignment with gaps (Boswell and McLachlan, 1984; Needleman and Wunsch, 1970), sequences are aligned locally with penalties for insertions and deletions, but the scores are given weights which die away exponentially with distance from the centre of each window. This newer method is more suitable for dealing with gaps, but is less susceptible to statistical analysis.

In the comparison matrices the highest observed score with a window of 30 corresponded to a double matching probability for the two peptide segments of only 0.53×10^{-10} (McLachlan, 1971) and was highly significant. We also calculated the highest expected score for the comparison of two random 344-residue sequences, i.e., the score which would be achieved on average just once with our protein. This score was exceeded not once, but 359 times in the natural sequence and showed that there are many significant repetitions.

In the damped Needleman-Wunsch method we used an effective range of 30 residues with rather low gap penalties: 2.0 to start each gap and 2.0 for each extension by one residue.

Acknowledgements

We thank Dr G.L.Everett (Johnson-Matthey Ltd., Royson) for determinations of the content of zinc and other metals, and Professor N.Hales and Dr G.Maguire (Department of Clinical Biochemistry, Cambridge University) for zinc determinations. We are grateful to Drs D.D.Brown and H.Pelham for advice, and our colleagues Drs D.Rhodes and H-C.Thøgersen for their help.

References

Adman,E.T., Sieker,L.C. and Jensen,L.H. (1973) *J. Biol. Chem.*, **248**, 3987-3996.

Bieker,J.J., Martin,P.L. and Roeder,R.G. (1985) *Cell*, **40**, 119-127.

Bieker,J.J. and Roeder,R.G. (1984) *J. Biol. Chem.*, **259**, 6158-6164.

Bogenhagen,D.F., Sakonju,S. and Brown,D.D. (1980) *Cell*, **19**, 27-35.

Bogenhagen,D.F., Wormington,W.M. and Brown,D.D. (1982) *Cell*, **28**, 413-421.

Bolognesi,M., Gatti,G., Menegatti,E., Guarneri,M., Marquart,M., Papamokos, E. and Huber,R. (1982) *J. Mol. Biol.*, **162**, 839-868.

Boswell,D.R. and McLachlan,A.D. (1984) *Nucleic Acids Res.*, **12**, 457-464.

Boulanger,Y., Goodman,C.M., Forte,C.P., Fesik,S.W. and Armitage,J.M. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 1501-1505.

Brown,D.D. (1984) *Cell*, **37**, 359-365.

Cornell,N.W. and Crivaro,K.E. (1972) *Anal. Biochem.*, **47**, 203-208.

Dayhoff,M.O. (1978) *Atlas of Protein Sequence and Structure, Vol. 5, suppl. 3*, published by National Biomedical Research Foundation, Washington D.C.

Denis,H. and le Maire,M. (1983) in Roodyn,D.B. (ed.), *Subcellular Biochemistry*, Vol. **9**, Plenum Publishing Co., pp. 263-297.

Engelke,D.R., Ng,S.Y., Shastry,B.S. and Roeder,R.G. (1980) *Cell*, **19**, 717-728.

Fersht,A. (1977) *Enzyme Structure and Mechanism*, published by W.H.Freeman & Co., San Francisco.

Ginsberg,A.M., King,B.O. and Roeder,R.G. (1984) *Cell*, **39**, 479-489.

Gottesfeld,J. and Bloomer,L.S. (1982) *Cell*, **28**, 781-791.

Hanas,J.S., Hazuda,D.J., Bogenhagen,D.F., Wu,F.Y.H. and Wu,C.W. (1983) *J. Biol. Chem.*, **258**, 14 120-14 125.

Honda,B.M. and Roeder,R.G. (1980) *Cell*, **22**, 119-126.

Hood,L.M., Campbell,J.H. and Elgin,S.C.R. (1975) *Annu. Rev. Genet.*, **9**, 305-334.

Laemmli,U.K.(1970) *Nature*, **227**, 680-685.

Laskowski,M., Kato,I., Kohr,W.J., March,C.J. and Bodard,W.C. (1980) in Peeters,H. (ed.), *Protides of Biological Fluids*, Vol. **28**, Pergamon Press, Oxford, pp. 123-128.

Lassar,A.B., Martin,P.L. and Roeder,R.G. (1983) *Science*, **222**, 740-748.

Lewis,C.D. and Laemmli,U.K. (1982) *Cell*, **29**, 171-181.

McLachlan,A.D. (1971) *J. Mol. Biol.*, **61**, 409-421.

McLachlan,A.D. (1980) in Jaenicke,R. (ed.), *Protein Folding*, Elsevier/North Holland, Amsterdam, pp. 79-96.

McLachlan,A.D. (1983) *J. Mol. Biol.*, **169**, 15-30.

Marshall,R.D. (1972) *Annu. Rev. Biochem.*, **41**, 673-702.

Moews,P.C. and Kretsinger,R.H. (1975) *J. Mol. Biol.*, **91**, 201-225.

Needleman,S.B. and Wunsch,C.D. (1970) *J. Mol. Biol.*, **48**, 443-453.

Ohlendorf,D.H. and Matthews,B.W. (1983) *Annu. Rev. Biophys. Bioeng.*, **12**, 259-284.

Pelham,H.R.B. and Brown,D.D. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 4170-4174.

Picard,B. and Wegnez,M. (1979) *Proc. Natl. Acad. Sci. USA*, **76**, 241-245.

Rossmann,M.G. and Argos,P.W. (1981) *Annu. Rev. Biochem.*, **50**, 497-532.

Sakonju,S., Bogenhagen,D.F. and Brown,D.D. (1980) *Cell*, **19**, 13-25.

Segall,J., Matsui,T. and Roeder,R.G. (1980) *J. Biol. Chem.*, **255**, 11986-11991.

Smith,D.R., Jackson,T.J. and Brown,D.D. (1984) *Cell*, **37**, 645-652.

Sottrup-Jensen,L., Claeys,H., Zajdel,M., Petersen,T.E. and Magnusson,S. (1978) *Prog. Chem. Fibrinolysis Thrombolysis*, **3**, 191-209.

West,T.S. (1969) *Complexometry with EDTA and Related Reagents*, published by BDH Chemicals Ltd., Poole, UK.

Received on 4 April 1985