The primary structure of transcription factor TFIIIA has 12 consecutive repeats

Raymond S. Brown, Christian Sander+ and Patrick Argos*

European Molecular Biology Laboratory, Postfach 10 2209 and *+Department of Biophysics, Max-Planck-Institut fur medizinische Forschung, D-6900 Heidelberg, FRG*

Received 22 April 1985

Analysis of the amino acid sequence of transcription factor TFIIIA from *Xenopus laevis* reveals the presence of 12 repeating structures, each about 30 residues in length These segments have been aligned and their secondary structure predicted The repeats each contain two invariant cysteines and two invariant histidines, perhaps to coordinate a zinc cation Possible nucleic acid interaction modes are discussed

Transcription factor *TFIIIA* *Primary structural repeat*

1. INTRODUCTION

Transcription factor TFIIIA is specifically required for the initiation of 5 S RNA synthesis in vitro by RNA polymerase III [1,2]. In previtellogenic oocytes of *Xenopus laevis*, the factor occurs in relative abundance in its association with 5 S RNA in a 7 S complex [3]. The factor also binds to a DNA sequence of about 50 base pairs in the intragenic control region as judged by DNase I protection experiments [4] and chemical modification studies [5]. TFIIIA is thus an example of a protein which binds specifically to both DNA and RNA.

The amino acid sequence of TFIIIA has been deduced recently from the nucleotide sequence of a cDNA clone [6]. An interesting region of homology was found between the intragenic control region of oocyte 5 S DNA and the TFIIIA gene which may indicate that both genes are under similar regulatory control. The authors further report that there appears to be little sequence homology between TFIIIA and other known DNA binding proteins Here, segments of the TFIIIA amino acid sequence have been aligned to reveal the presence of internal repeating structures

2. METHODS

The TFIIIA sequence was compared with itself to check for repeating, structurally homologous segments Every possible pairwise comparison of sequence spans of 20 residues in length was assessed by 2 scoring procedures: the Dayhoff relatedness odds matrix [7–9] whose elements express relative weights with which amino acids substitute for one another in aligned sequences in 71 protein families, and calculation of the mean correlation coefficient over 6 residue physical characteristics [10,11] important for protein folding (cf. [12]). The characteristics (α-helix, β-sheet, and turn secondary structural conformational preferences; residue polarity; and 2 hydrophobicity measures) are listed by Argos et al. [10]. The 2 scores were then scaled and combined. When the TFIIIA search matrix was complete over all 20-residue matches, all coefficients were recalculated as a number of standard deviations (σ) above the mean matrix coefficient These fractional standard deviations were then placed in the

* On sabbatical leave from Purdue University, Department of Biological Sciences, West Lafayette, IN 47907, USA

Published by Elsevier Science Publishers B V (Biomedical Division)
00145793/85/$3 30

matrix corresponding to the beginning residue number of the 2 oligopeptides compared. The comparison method has been described in detail in [11].

A more sensitive approach to delineation of the TFIIIA repeat locations involved a summation of the matrix values along a given line at each TFIIIA sequence position, excluding those of the exact self-comparison and including those above a certain threshold value (3.0 σ) The sums were then divided by the expected number of repeats minus the self-repeat (e.g. TFIIIA, 8) to prevent large, isolated values along a line from dominating the results If the number of values above the threshold was greater than the number of expected peaks, the larger number was used as a divisor to prevent overlapping matrix values near repeat termini from influencing the search. Finally a linear plot of the TFIIIA sequence position vs the averaged sum of the standard deviation fractions was used to delineate the beginning residue of each repeat. The starting sequence position for the repeating units would be expected to occur near points where the averaged sum increased dramatically This latter plot was smoothed by a sliding averaging procedure [13] over 3 successive points and for 10 complete cycles for easier visual observation of any repeats.

Plots of the sequence number vs the conformational preference parameter (α-helix, β-strand, turn [14]) for a given amino acid were determined for each protein region using a least-squares smoothing procedure The smoothing process was repeated for 3 cycles over each of the parametric plots The smoothed curves for each potential were then averaged over the aligned TFIIIA sequence repeats, a procedure which should yield a better prediction than that from any one sequence [15] The rules used to assign secondary structural types at each residue position have been reported in [11].

3. RESULTS AND DISCUSSION

The similarity matrix shown in fig.1 reveals that a sequence element of approx. 30 residues is repeated. A summation of matrix values along a given line at each TFIIIA sequence position was divided by 8. Only values above 3.0 σ were used. The normalized sums were then plotted against sequence number and the curve smoothed over 10 cycles (fig.2). The plot shows at least 9 and probably 10 repeats

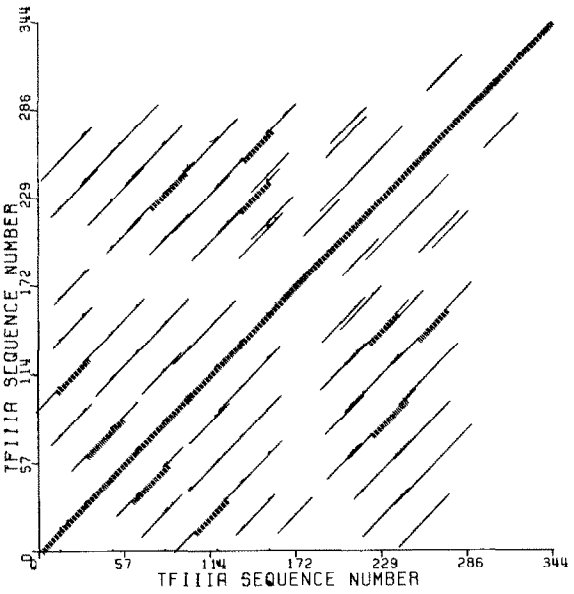

Fig 1 Self-comparison homology search matrix for TFIIIA The search window length used was 20 residues Lines are shown for all search values between 3 0 and 5 0 σ while vertical bars indicate peaks greater than 5 0 σ The symbols are plotted over the entire probe length with the higher peaks allowed to dominate for any overlapping search values The self-search is necessarily symmetric about the strong main diagonal resulting from the self-comparisons The obvious 9 dominant lines point to the strongly repeating units in TFIIIA

An optimal alignment of the repeats was achieved by placing the cysteine and histidine residues in register and then matching the remaining residues to obtain the best relationship amongst nucleic acid codons with regard to the hydrophobic and polar character of the amino acids. This allowed a satisfactory alignment of the sequence from residues 1–276 as shown in fig.3 The addition of residues 277–304 was possible because of the strong homology between residues 144–152 and 291–298. Similarly the segment 305–326 was aligned owing to a homology between residues 166–180 and 309–322. Finally the remaining C-terminal fragment 327–344 was added on the basis of the strong sequence similarity between 314–323 and 328–337

A number of near invariant amino acids are present in the aligned repeated sequences. These are

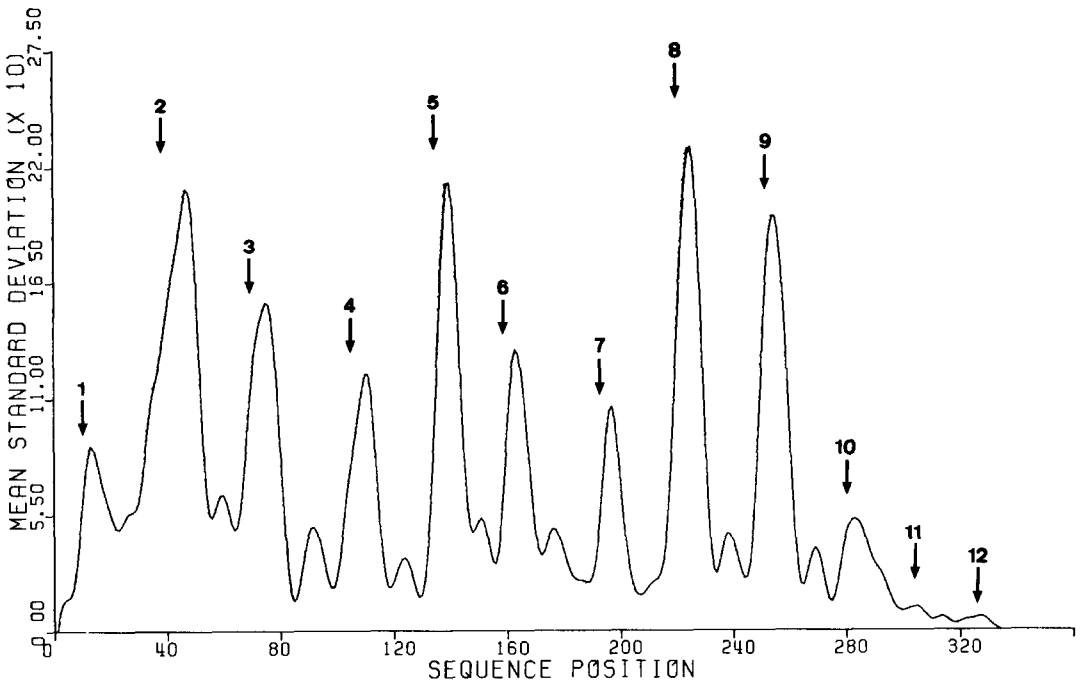

Fig 2 Plot of the TFIIIA sequence position against the averaged sum of the search matrix elements with value 3 0 σ or greater In this case, as juxtaposed to the fig.1 matrix, the search values were placed only at the beginning residue numbers for each oligopeptide match The best results were obtained with a search window of 5 residues Arrows indicate the starting sequence position of each repeat The repeat starts are obviously at plot points where the summed matrix values show a sharp increase The plot has been smoothed over 10 cycles by a sliding-window averaging technique based on 3 successive points The curve indicates clearly 10 of the 12 repeats

```
Repeat
          1        10        20        30        39
1     (1) MGEKALPVVYKRYICSFAD-CGAAYNKNWKLQAHLCK-H
2    (38) TGEK--P--F---PC-KEEGCEKGFTSLHHLTRHSL-TH
3    (68) TGEKN----F---TC-DSDGCDLRFTTKANMKKHFNRFH
4    (99) NI-K--ICVY--V-CHFEN-CGKAFKKHNQLKVH-QFSH
5   (130) T-Q-QLP--Y-E--CPHE-GCDKRFSLPSRLKRH-EKVH
6   (160) AG-------Y---PCKKDDSCSFVGKTWTLYLKHVAECH
7   (189) --Q-DLA-V-----C---DVCNRKFRHKDYLRDH-QKTH
8   (215) --EKER-TVY---LCPR-DGCDRSYTTAFNLRSHIQSFH
9   (247) -EE-QRP--F--V-CEHA-GCGKCFAMKKSLERHSV-VH
10  (277) DPEK-R-KL-KE-KCPR--P--KR-SLASRLTGYIP--P
11  (305) ---K----S-KE----KNASVSGTEKTDS-LVK--NK--
12  (327) --------------P-----SGTETNGS-LVLDKLTIQ
        ++      +   *   + * + +     +   *     *
        tttccceßßßßccßßßttttttttttaaaaaaaaaaccccccc
```

Fig 3 Alignment of the repeat structure segments in TFIIIA The start residue number is given in parentheses for each repeat. The alignment position numbers referred to in the text are also given (1–39) Mean secondary structure predictions α, α-helix, β, β-strand; t, turn; c, coil The conserved cysteines and histidines are indicated by a * while conservation at a particular alignment position is indicated by a + if 7 or more of the residues displayed conservative substitution according to the following scheme· (P,G); (T,S), (Q,E,D,N), (A,V,I,L,M,C,A,F,Y,W)

phenylalanine or tyrosine at alignment positions 10 and 25, cysteine at positions 15 and 21, leucine at position 31, histidine at positions 34 and 39, and an acidic residue at positions 18 or 19. This pattern degenerates in the 3 C-terminal repeats although leucine is still found at position 31. The invariant cysteine residues are separated by up to 5 other amino acids.

There is no published evidence at present in support of the existence of disulphide bonds within TFIIIA. There are chemical experiments suggesting at least 2 Zn^{2+} binding sites [16] Since cysteine and histidine are often observed zinc ligands in proteins and since the cation is generally tetrahedrally coordinated in several known tertiary protein structures [17], it is proposed that each of the 9 TFIIIA repeating units, containing 2 cysteines and 2 histidines, may coordinate one zinc ion The mean secondary structure prediction supports this proposal as the residues between the cysteine and histidine pairs are predicted in a turn or

coil configuration, allowing the main chain to loop back upon itself for zinc coordination. Apparently the 3 C-terminal repeats have lost this ability as 4 cysteines and histidines are not present Nonetheless, their close homology would point to their involvement in the possible gene duplicating events leading to the present-day TFIIIA molecule.

The National Biomedical Research Foundation data bank of protein sequences [18], consisting of 526120 amino acids distributed over 2675 proteins, was searched for the following pattern displayed by the N-terminal 9 repeats cysteine, any 2–5 amino acids, cysteine, any 12 residues, histidine, any 3–4 residues, and histidine. No protein sequences displaying this pattern were found in the present bank

It has been proposed that, from the known 3-dimensional structures of various bacterial DNA repressor proteins, a helical structure interacts with the major groove of double-stranded DNA [19,20] An α-helix is predicted for TFIIIA alignment positions 25–32 (fig.3). Within the helices as well as bordering their termini are several basic residues. It has been suggested that TFIIIA binds in an extended fashion [21] to the 50 base pair intragenic control region of the 5 S RNA gene It is thus proposed here that the repeating helical segments could interact with DNA in an extended mode. The plethora of basic residues may bind DNA phosphates, providing a molecular anchor for TFIIIA It is noteworthy that the most conserved residues in addition to cysteines and histidines are contained within the predicted helical span; namely, phenylalanine or tyrosine at alignment position 25 and leucine at position 31, residues which may be significant for the association of TFIIIA with DNA. The Zn^{2+} centres with their coordinated cysteine and histidine side chains might provide structural stability to each of the local units associating with the extended DNA The N-terminal regions of the repeating units, which contain many predicted turn and coil residues, could provide sufficient flexibility to allow the stabilized local core units to encompass the nucleic acid. Though this structural model is clearly speculative, the observations made here suggest biochemical experiments that can be performed on TFIIIA such as site-directed mutagenesis and identification of the cation binding sites

ACKNOWLEDGEMENTS

We thank Dr D.D. Brown for communicating the incomplete cDNA sequence data. Ms Ines Benner of EMBL was essential in the preparation of the manuscript.

REFERENCES

[1] Bieker, J.J , Martin, P L and Roeder, R G (1985) Cell 40, 119–127
[2] Setzer, D R and Brown, D D (1985) J Biol Chem 260, 2483–2492
[3] Picard, B and Wegnez, M (1979) Proc Natl Acad Sci USA 76, 241–245
[4] Engelke, D R , Ng, S -Y , Shastry, B S and Roeder, R G (1980) Cell 19, 717–728
[5] Sakonju, S and Brown, D D (1982) Cell 31, 395–405
[6] Ginsberg, A M , King, O B and Roeder, R G (1984) Cell 39, 479–489
[7] Barker, W C , Ketcham, L K and Dayhoff, M O (1978) in Atlas of Protein Sequence and Structure (Dayhoff, M O ed) vol 5, Suppl 3, pp 359–362, National Biomedical Research Foundation, Washington, DC
[8] McLachlan, A D (1971) J Mol Biol 61, 409–425
[9] Staden, R (1982) Nucleic Acids Res 10, 2951–2961
[10] Argos, P , Hanei, M , Wilson, J M and Kelley, W N (1983) J Biol Chem 258, 6450–6457
[11] Zalkin, H , Argos, P , Narayana, S V L , Tiedeman, A A and Smith, J M (1985) J Biol Chem , in press
[12] Creighton, T E (1978) Biophys Mol Biol 33, 231–297
[13] Kyte, J and Doolittle, R F (1982) J Mol Biol 157, 105–132
[14] Palau, J , Argos, P and Puigdomenech, P (1982) Int J Protein Peptide Res 19, 394–401
[15] Argos, P , Schwarz, J and Schwarz, J (1976) Biochim Biophys Acta 439, 261–273
[16] Hanas, J S , Hazuda, D J , Bogenhagen, D F , Wu, F Y -H and Wu, C -W (1983) J Biol Chem 258, 14120–14125
[17] Argos, P , Garavito, R M , Eventoff, W , Rossmann, M G and Branden, C I (1978) J Mol Biol 126, 141–158
[18] National Biomedical Research Foundation (1984) Amino Acid Sequence Data Bank, Georgetown University, Silver Springs, MD
[19] Ohlendorf, D H , Anderson, W F , Lewis, M , Pabo, C O and Matthews, B.W (1983) J Mol Biol 169, 757–769
[20] Pabo, C O and Sauer, R T (1984) Annu Rev Biochem 53, 293–331
[21] Smith, D R , Jackson, I J and Brown, D D (1984) Cell 37, 645–652