The human EBNA-2 coactivator p100: multidomain organization and relationship to the staphylococcal nuclease fold and to the tudor protein involved in *Drosophila melanogaster* development

Isabelle CALLEBAUT* and Jean Paul MORNON

Systèmes Moléculaires et Biologie Structurale, Laboratoire de Minéralogie-Cristallographie, CNRS URA09, Universités Paris 6 & Paris 7, case 115, 4 place Jussieu, 75252 Paris Cedex 05, France

The human p100 protein was recently identified as a coactivator of the Epstein-Barr virus nuclear antigen 2. On the basis of sequence analysis with the hydrophobic cluster analysis method, we predict that this protein consists of a repeat of four similar domains. Their fold can be related to the staphylococcal nuclease structure whose first subdomain belongs to the large oligonucleotide/oligosaccharide-binding (OB)-fold superfamily. The catalytic amino acids present in nucleases are missing, however, suggesting that these repeated motifs could only serve to bind DNA without catalytic activity, as in many other OB-

folds. A highly modified fifth domain follows the four nuclease-like domains, conserving the second subdomain of the nuclease structure but not the first one (the OB-fold), which is replaced by an original domain found in multiple copies in the tudor protein, a *Drosophila melanogaster* protein required during oogenesis for establishment of a functional posterior organizing centre. We named this heretofore undescribed domain the ‘tudor domain’ and highlight within it five invariant residues which could be involved in one of the essential roles played by these proteins.

INTRODUCTION

A novel nuclear protein, p100, was recently found to bind to the Epstein-Barr virus nuclear antigen 2 (EBNA-2) and coactivate gene expression mediated by the EBNA-2 acidic domain [1]. This coactivating effect was suggested to be due to the interaction of p100 with both the p56 and p34 subunits of TFIIE, an essential component of the RNA polII initiation complex [2,3]. Hence, p100 could fulfil its role of coactivator by serving as a bridge between the EBNA-2 acidic domain and the basal transcription machinery. Consistent with its role as a coactivator, p100 is

essential for cell viability and appears to be developmentally regulated. On the other hand, p100, like another coactivator PC4, is also able to bind single-stranded DNA [1]. This protein could therefore serve several functions and be related to other localized nuclear complexes, such as splicing and replication complexes.

In order to further elucidate the structural and functional features of this protein, which has been reported to lack significant similarity with any other protein (as deduced from classical sequence analysis), we have used hydrophobic cluster analysis (HCA) [4,5]. This powerful method is indeed able to detect three-dimensional similarities between proteins showing very limited sequence relatedness. Its sensitivity at low levels of sequence identity (typically below 30%) stems from its ability to significantly detect secondary structure elements [6]. The effectiveness of the HCA method has been widely demonstrated, for instance, in detecting catalytic residues in families of divergent enzymes [7,8], in identifying repeated motifs in several proteins [9,10] and, more recently, in predicting one of the functions of an important molecular chaperone [11], at similar levels of sequence identity to those encountered in the present study.

Here we predict that p100 is composed of a repeat of four similar domains sharing between 20 and 33% sequence identity whose structure can be related to the staphylococcal nuclease (SN)-fold [12–14]. SNs (EC 3.1.31.1), also called thermonucleases, are small calcium-dependent enzymes which hydrolyse both DNA and RNA, yielding 3'-mononucleotides and polynucleotides. They form a rather well conserved family of proteins including SNs [15], a nuclease from the pSa plasmid of *Shigella flexneri* [16], the parB protein of *Escherichia coli* RP4 [17] and a hypothetical protein (OrfA) from plasmid R100 [18]. The sequences of these enzymes contain a conserved motif described in the PROSITE databank [19] that includes two active site

Figure 1 Ribbon diagram of the structure of the *S. aureus* nuclease (PDB identifier: 1SYG) obtained using MOLSCRIPT software [20]

The OB-fold (subdomain A, not shaded, strands $\beta 1$ to $\beta 5$) is highlighted. The two additional α -helices (subdomain B, helices $\alpha 2$ and $\alpha 3$) are shown shaded. The loops L12, L3 α 1 and L45, designated according to the regular secondary structures they connect, contribute residues to the substrate-binding sites of OB-folds [21].

Abbreviations used: HCA, hydrophobic cluster analysis; SN, staphylococcal nuclease; OB, oligonucleotide/oligosaccharide binding; EBNA-2, Epstein-Barr virus nuclear antigen 2.

* To whom correspondence should be addressed.

Table 1 OB-fold proteins whose structures have been solved

ON and OS indicate oligonucleotide and oligosaccharide-binding activities respectively. * Cold shock protein is related to the large group of 'Y-box' nucleic acid binding proteins [23]. Proteins from the same family are indicated within parentheses after the first family member.

Protein	References	Activity
Staphylococcal nuclease	[12–14]	ON
Aspartyl t-RNA synthetase	[24]	ON
(Lysyl t-RNA synthetase)	[25]	
Cold shock protein*	[26]	ON
Gene V protein	[27]	ON
Ribosomal S17	[28]	ON
Heat-labile enterotoxin	[29]	OS
(Cholera toxin)	[30]	
Verotoxin	[31]	OS
N-terminal domain of	[32]	?
<i>Staphylococcus enterotoxin B</i>		
(Toxic shock syndrome toxin 1)	[33]	
Active domain of tissue inhibitor of	[34]	?
metalloproteinase 2		
<i>Aspergillus giganteus</i> antifungal protein	[35]	?

residues, an arginine and a glutamic acid. However, these amino acids, as well as several others involved in the catalytic mechanism, are missing in the p100 SN-like domains described in this report, suggesting that these domains could be limited to DNA binding without catalytic activity. This 'basic' DNA-binding activity suggested here for the p100 SN-like domains is further supported by the fact that the first subdomain (98 amino acids) of the structure of the *Staphylococcus aureus* nuclease adopts an oligonucleotide/oligosaccharide-binding (OB)-fold (Figure 1), which has been observed in several proteins having in common the property of binding oligonucleotides or oligosaccharides, without necessarily having an enzymic activity ([21,22]; Table 1). The OB-fold consists of a five-stranded Greek-key β -barrel often capped by an α -helix located between the third and fourth strands (Figure 1). All OB-fold proteins share the same β -sheet topology, suggesting a common ancestry in their evolution, despite the lack of sequence similarities between them [21]. The binding sites for oligonucleotides or oligosaccharides, when known, are found on the same part of the barrel surface and involve in particular three variable loops (L12, L3 α and L45; Figure 1) [21].

We discuss here the implications of the structural features of p100 for its mechanism of action, especially the potential DNA interacting regions found by similarity with the well-described OB-fold. In addition, we suggest that the p100 C-terminal domain, which follows the four SN-like domains, could consist of a 'hybrid' SN-like domain where the first subdomain (the OB-fold) is completely different and is found in multiple copies in the *Drosophila melanogaster* tudor protein. Within this domain, invariant amino acids, which may play an essential role in the function of p100 and the tudor protein, were identified.

EXPERIMENTAL

Systematic searches of databanks [36,37] allow detection of sequences which could belong to the same functional and/or structural family. However, at the low levels of sequence identity (25–30%) often observed, these automatic methods are unable to distinguish similarities due to structural relationships from background noise. The HCA method is helpful in this regard in so far as it allows comparison of not only the sequences but also

Figure 2 Correction introduced in the N-terminal end of the human p100 sequence

In the HCA plots represented here, the protein sequences are written on a duplicated α -helical net and the contour of clusters of hydrophobic residues are automatically drawn. The α -helical net has been demonstrated to offer the best correspondence between the positions of hydrophobic clusters and regular secondary structures [6]. The standard one-letter code for amino acids is used except for proline (regular secondary structure breaker), glycine (the less constraint amino acid), serine and threonine (which can be accommodated either at the protein surface or in a hydrophobic environment masking the hydroxyl group through H-bonding with the polypeptide backbone) which are represented by the symbols ★, ◆, ◻ and ◻ respectively. The first hydrophobic cluster of the human p100 sequence, designated as *a*, seems to be incomplete when compared with the equivalent motif *a* in the *C. elegans* sequence, as well as to motifs *a* in similar domains (see Figure 3). Translation of the cDNA sequence upstream of the predicted start codon (Met-1) led to a sequence harbouring the main characteristics of motif *a* (human 'revised').

the protein secondary structures statistically centered on hydrophobic clusters, as well as their distribution [6]. Similar plots could therefore indicate similar three-dimensional folds. Guidelines to the use of this method are given in [4,5].

The HCA score is proportional to the hydrophobic amino acids which are topologically conserved (often not chemically identical), and therefore reflects the degree of conservation of the hydrophobic core. High HCA scores are associated with low root mean squares values between three-dimensional structures [5].

The accuracy of the alignments can be assessed by computing identity or similarity Z scores; these represent differences between the considered alignment identity or similarity score and the mean score of a distribution computed for alignment of sequence 1 versus a large number of random shuffled versions of

sequence 2 (10000 in the present study). These differences are expressed relative to the S.D. of the random distribution.

RESULTS

Detection of a repeated domain in the p100 sequence

Recently, a protein identified from sequencing of *Caenorhabditis elegans* chromosome III (genpep identifier: celf10g79 [38]) was

found to be similar to the human p100 sequence (genpep identifier: hsu22055), sharing with it 49.3% sequence identity over the entire protein length (913 and 883 amino acids for *C. elegans* and human sequences respectively). In comparison with the human p100, the *C. elegans* p100 appears to possess an additional N-terminal sequence of about 30 amino acids which mainly contains hydrophilic residues (Figure 2). However, the absence of such a domain in the human p100 sequence seems to

Figure 3 HCA plots of the human p100-repeated domains (p100-I to p100-V) and the *S. aureus* nuclease (SNa) whose three-dimensional structure is known [12–14]

SNa amino acid numbering is reported conventionally, with the main residues involved in the catalytic site shown with arrows (D21, R35, D40, E43, K84, Y85 and R87). SNa secondary structures as well as its division into two subdomains A and B are shown above its HCA plot. Nine amino acids were added to the N-terminal of the reported human p100 sequence [1] as described in Figure 2. A similar distribution of hydrophobic clusters with similar characteristics (designated as motifs a to h) can be deciphered all along the defined domains. The vertical lines indicate the proposed correspondences between sequences; clusters with very similar characteristics are darkly shaded whereas sequence identities are indicated in italics within open circles. However, motifs a–e are not found, as such, in the fifth p100 domain, which led us to differentiate the sequence linking the p100-IV and p100-V subdomains B and to name it subdomain A' (boxed).

Figure 4 HCA-based alignment of the sequences of the four p100 repeated domains (numbered from I to IV) and the *S. aureus* and *Staphylococcus hyicus* nucleases (SNa and SNh respectively). Subdomain B of a fifth p100 domain (V) is aligned relative to the subdomains B of the first four p100 domains

Secondary structures of SNa are shown above the sequences. *t*, *m* and *b* are positions in the barrel top, middle and bottom layers respectively, as described in [21]. The middle and bottom layers are mainly composed of hydrophobic or neutral residues (mainly β -branched residues for the bottom layer). White letters within black boxes indicate amino acids that are identical between SN and p100, and bold letters within white boxes those which are identical only between the p100-repeated domains. Residues which are proposed to be involved in the catalytic mechanism of SNs are indicated with the symbol #, and conserved hydrophobic and aromatic amino acids with the symbols o and Φ respectively.

be an artefact due to an incorrect choice of its start codon, as deduced from subsequent analysis (see the following section). The human p100 sequence would consequently contain at least nine additional amino acids (PTVQRGIHK) which show significant similarity with the *C. elegans* p100 sequence (Figure 2).

Examination of the HCA plots of the human and *C. elegans* p100 sequences reveals the presence of similar hydrophobic motifs at regular intervals (about every 150 amino acids), suggesting the possibility of a 4-fold repeated domain (domains I to IV; Figure 3). Two hydrophobic clusters are especially well conserved, designated as *d* and *f* in Figure 3, but other anchor points can be found all along the domains defined in this manner. In particular, the repeated domains contain at their ends (motif *h*) a conserved tryptophan (except for domain III where it is probably replaced by a histidine) after a rather hydrophilic region (Figures 3 and 4). Furthermore, in the first part of the repeated domain, a proline (in domain I, a glycine) is always found three residues after an invariant isoleucine ending motif *c* (Figures 3 and 4). All the sequences linking motifs *a* and *b* contain a glycine (except for an alanine in domain IV), which probably plays an important role in a loop structure.

Curiously, the *f* and *g* motifs as well as the conserved tryptophan in motif *h* can be found once again in the end of the human and *C. elegans* sequences (subdomain 'B' of the fifth domain in Figure 3), but the motifs constituting the N-terminal

two thirds of the domains I, II, III and IV (subdomains 'A') are not found upstream of the *f* motif. In addition, the distance between the *f* motifs of the fourth and fifth domains (about 240 amino acids) is greater than those observed between the motifs of the four first domains (about 150 amino acids). These two observations suggest that the first part of the fifth p100 domain, designated as subdomain A', differs largely or is totally different from the subdomains A of the first four domains.

Table 2 gives the sequence identity, sequence similarity and HCA score values calculated between the repeated domains (with only subdomain B for domain V). The high HCA score values reflect overall conservation of the hydrophobic core and the regular secondary structures constituting it. An objective way to assess the validity of predictions is to calculate Z-scores from the comparison of a real alignment with a large number of alignments performed after randomization of one of the compared sequences. These values, calculated for the sequence identity, sequence similarity and HCA scores, are also reported in Table 2 and allow evaluation of how the signal emerges from background. Confidence is assumed for values exceeding 6, whereas values above 3 are often encountered for similar three-dimensional structures sharing low-level sequence identity [39].

Sequence identity values between the p100 repeated domains range between 19.8 and 32.6% (mean 22.7%) and corresponding Z-scores between the entire domains are observed to be between 6.3 and 13.2 (mean 8.7). These results together with the cor-

responding good values for similarity and HCA scores, as well as the good ratio between the observed Z3 index (the product of the three Z scores) and best random Z3 (Table 2), clearly confirm the structural relationship of the p100 domains.

It is worth noting that domain I is more similar to domain III than to domain II, which itself shares more similarity with domain IV, suggesting that p100 could have evolved from the duplication of a previously duplicated motif.

The fold of the p100 repeated domains has similarity with that of SN

Computer searches of sequence databanks with each of the defined 150 amino acid domains, or with profiles deduced from the multiple alignment shown in Figure 4, reveal sequence and structural similarities with SNs (Figures 3 and 4). The sequence identity, sequence similarity and HCA score values are indicated in Table 3 and were found to be significant or highly significant, as assessed by the corresponding Z scores. It therefore appears likely that the p100 repeated domains and SN, whose three-dimensional structure is known [12–14], have a similar fold, consisting of a highly twisted five stranded β -barrel and three helices (Figure 1).

The two subdomains consisting of the 'OB-fold' ($\beta 1$, $\beta 2$, $\beta 3$, $\alpha 1$, $\beta 4$ and $\beta 5$) and the two C-terminal helices ($\alpha 2$ and $\alpha 3$) are all retrieved well conserved in the p100 nuclease-like domains I, II, III and IV (Figures 3 and 4). It is worth noting that the $\beta 1$ strand corresponds to a cluster of considerable length (since it belongs to both β -sheets, Figure 1) in all nuclease-like domains with the exception, in a first analysis, of the first human p100 domain (Figures 2 and 3). This anomaly seems to be due to an incorrect choice of the start codon, since translation of the DNA sequence upstream of the Met-1 reconstituted a strand possessing all the $\beta 1$ structural characteristics (Figure 2).

Gaps between the p100 nuclease-like domains and SN sequences were found in loops, especially in the loops linking $\beta 2$ to $\beta 3$ (L23), $\beta 3$ to $\alpha 1$ (L3 $\alpha 1$), $\beta 4$ to $\beta 5$ (L45), $\beta 5$ to $\alpha 2$ (L5 $\alpha 2$) and $\alpha 2$ to $\alpha 3$ (L $\alpha 2\alpha 3$) (Figure 4).

The function of the four p100 nuclease-like domains would appear to differ from that of classical SNs, i.e. hydrolysis of phosphodiester bonds, as the amino acids proposed to be involved in catalysis are not conserved. This mechanism implicates a Ca^{2+} ion, a catalytic water acting as the attacking nucleophile, a Glu (Glu-43) and an Arg (Arg-87) residue, probably acting as the general base and general acid respectively. Several residues are involved in the positioning of the 5' and 3' phosphates (Arg-35, Arg-87, Lys-84, Tyr-85) and of the Ca^{2+} (Asp-21, Asp-40, Thr-41) [40]. Most of these residues, highlighted in Figures 3 and 4, are missing in the p100 repeated SN-like domains, suggesting that these SN-like domains, as well as other already known OB-folds (Table 1), may be limited to DNA binding. This 'basic' function has been shown experimentally for p100 [1].

The p100 domain V appears to correspond to a 'hybrid' structure, containing only the two C-terminal helices of the SN-fold (subdomain 'B') without the classical OB-fold preceding it. Whether subdomain 'A' of this domain V corresponds to a highly modified OB-fold supporting insertions of considerable length, or to a totally different structure, is difficult to predict, as the OB-fold can tolerate a wide variety of sequences showing little or no sequence identity [21]. Further investigations were carried out on this particular subdomain, whose sequence conservation between human and *C. elegans* is otherwise far lower than that observed between their nuclease-like domains (34.8% versus 53.1%).

Table 2 Sequence identity [i (%)], similarity (s) and HCA [h (%)] scores calculated between the human p100 repeated domains. Corresponding Z-scores (see Materials and methods section) are indicated in parentheses

The Z3 values (product of the three Z-scores) for the proposed alignments and the best Z3 value obtained for random calculations are shown in italics. Z-scores higher than 6 can be considered as highly significant whereas scores higher than 3 are frequently encountered for three-dimensionally-related structures with low levels of sequence identity. *Only the small subdomain (subdomain B) was considered for the calculations relative to domain V (see text for details).

		p100-II	p100-III	p100-IV	p100-V*
p100-I	i	23.7 (8.1)	32.6 (13.2)	19.8 (6.3)	28.6 (6.0)
	s	0.68 (7.4)	1.17 (10.5)	0.55 (6.9)	1.43 (6.1)
	h	64 (6.3)	63 (5.6)	68 (7.0)	61 (7.3)
	Z3	<i>378/36</i>	<i>837/30</i>	<i>304/38</i>	<i>271/68</i>
p100-II	i	—	20.0 (6.6)	29.8 (10.3)	25.6 (4.8)
	s	—	0.41 (5.9)	1.39 (10.0)	1.02 (4.5)
	h	—	58 (4.5)	67 (6.1)	55 (6.8)
	Z3	—	<i>177/41</i>	<i>631/49</i>	<i>148/37</i>
p100-III	i	—	—	23.0 (7.9)	22.2 (4.2)
	s	—	—	0.64 (7.0)	62 (3.7)
	h	—	—	62 (5.7)	62 (8.2)
	Z3	—	—	<i>317/46</i>	<i>130/30</i>
p100-IV	i	—	—	—	21.9 (3.4)
	s	—	—	—	1.05 (4.6)
	h	—	—	—	76 (10.6)
	Z3	—	—	—	<i>166/32</i>

Table 3 Sequence identity [i (%)], sequence similarity (s) and HCA [h (%)] scores calculated between the p100 repeated domains and the *S. aureus* (SNaureus) and *S. hyicus* (SNhyicus) nuclease sequences

These two nucleases share 41.7% sequence identity. Corresponding Z-scores (see Material and methods section) are indicated in parentheses. The Z3 values (product of the three Z-scores) for the proposed alignments and the best Z3 value obtained for random calculations are shown. Z-scores higher than 6 can be considered as highly significant, whereas scores higher than 3 are frequently encountered for three-dimensionally-related structures with low levels of sequence identity. *Only the small subdomain B was considered for the calculations relative to domain V (see text).

		SNaureus	SNhyicus
p100-I	i	19.5 (6.4)	22.9 (8.0)
	s	0.70 (7.3)	0.88 (8.2)
	h	56 (5.1)	60 (5.8)
	Z3	296/32	498/50
p100-II	i	20.8 (6.9)	18.0 (5.7)
	s	0.58 (6.4)	0.32 (5.2)
	h	56 (4.4)	43 (2.5)
	Z3	276/30	113/33
p100-III	i	20.7 (6.8)	18.9 (6.1)
	s	0.37 (5.3)	0.36 (5.3)
	h	47 (3.3)	49 (3.8)
	Z3	221/31	235/23
p100-IV	i	20.8 (6.2)	14.7 (4.0)
	s	0.66 (6.7)	0.34 (5.4)
	h	58 (4.9)	52 (4.0)
	Z3	358/40	160/40
p100-V*	i	23.8 (4.1)	26.8 (4.9)
	s	0.95 (4.2)	0.56 (3.2)
	h	57 (6.8)	52 (5.8)
	Z3	118/47	91/43

Subdomain A' of p100 domain V is found to be a repetitive element of the tudor protein and shares highly conserved motifs

The subdomains A' of the human and *C. elegans* p100 domain V were also searched in databanks for proteins sharing similar

Figure 5 Tudor domains within p100 and tudor proteins

Top: HCA comparison of 'tudor domains'. Subdomain A of the human p100 domain V and two of the eight tudor domains of the tudor protein. Similar clusters are darkly shaded whereas sequence identities are indicated in italics within open circles. A more conserved segment of about 80 amino acid acids, containing the motifs x, y and z described below, is boxed. Middle: schematic representation of the organization of the p100 and tudor proteins into domains. Bottom: sequence conservation within the tudor domains in the p100 and tudor proteins focused on motifs x, y and z. Invariant and highly conserved residues are shown in black and white boxes respectively. Conserved hydrophobic amino acids are indicated with stars. The two additional tudor domains (domains I' and II'; not shown) conserve the motifs y and z or motif z respectively, but show higher divergence in the rest of their sequences.

features. Interestingly, they show similarity (28.6 % identity over 83 amino acids) with a segment of the tudor protein (between amino acids 1800 and 1880), a *D. melanogaster* protein required in oogenesis for establishment of a functional centre of morphogenic activity in the posterior tip of the *Drosophila* embryo [41–43].

HCA analysis (Figure 5, top) of this large protein (2515 amino acids) revealed: (1) that the similarities can be extended on both sides of the highlighted region to the entire length of the p100 subdomain A' (about 200 amino acids); and (2) that the so-defined domain is repeated eight times along the tudor protein sequence (Figures 5, all three panels). The tudor-repeated domains and p100 subdomains A', which we suggest be designated by the generic term 'tudor domains', show a high degree of divergence (far below 20 % sequence identity on the entire length of the domain), but all include a segment of about 80 amino acids which can be aligned, containing three highly conserved motifs ($x = \text{ooo}[\text{QH}]$, $y = \text{D} \times \times [\text{WFYL}]\text{YR}[\text{ACS}]$ and $z = \text{ooooD}[\text{YF}]\text{G}[\text{NY}]$, where o and $\times$ are hydrophobic and any amino acid respectively) contain five invariant residues: two Asp, one Tyr, one Arg and one Gly (Figure 5). They correspond to Asp-716, Tyr-720, Arg-721, Asp-739 and Gly-741 in the human p100 sequence. The second Asp is found after a stretch of hydrophobic amino acids (motif z) which would correspond to a β -strand. In addition, a Gln or His is always found in motif x , also at the end of a probable β -strand.

The polar and invariant character of four of these residues suggests the potential involvement of the tudor domains in a hydrolytic mechanism. However, the sequence signature deduced from the multiple alignment of motifs y and z ($\text{D} - \times (2) - [\text{WFYL}] - \text{Y} - \text{R} - \times (10,14) - [\text{VIL}] - \times (2) - [\text{VIL}] - \text{D} - [\text{FY}] - \text{G} - [\text{NY}]$; the invariant residues are in bold) is found to be specific to the tudor and p100 proteins when searched against the Swiss-Prot and Genpep databases and does not allow them to be linked to an already described family of enzymes.

DISCUSSION

Many proteins share structural and functional similarities as a result of a common evolutionary origin. However, this common origin is often obscured by the extent of sequence divergence and can only be deciphered after experimental resolution of the protein structures [44]. In this context, the present study illustrates the efficiency of the HCA method in overcoming the limitations associated with classical sequence analysis in detecting structural similarity between protein sequences or within them (repeated motifs) at low levels of sequence identity (typically below 30 %). For instance, a duplicated immunoglobulin-like domain was successfully detected by this method in the extracellular domain of cytokine receptors at a very low level of sequence identity ($\sim 10\%$) [10]. We report here the presence of repeated domains sharing low levels of sequence identity within the human EBNA-2 coactivator p100 and in the *D. melanogaster* tudor protein, both repeats which had gone undetected with classical methods of sequence analysis [1,42].

The tudor protein is one of the maternally-encoded gene products that contributes to proper functioning of the *Drosophila* posterior plasm [41]. Here we show that this protein is composed of eight similar domains (the so-called 'tudor' domains). The presence of two additional but more divergent tudor domains can be hypothesized at the beginning of the sequence (Figure 5, middle, domain I') and between the hinge following domain II and domain III (Figure 5, middle, domain II'). These two

domains possess some of the motifs characterizing the tudor domain (y and z in domain I', and z in domain II'), but both would be largely modified since the other motifs, in particular motif x , and other sequence characteristics deduced from HCA are not found, unlike in the eight well-defined tudor domains of the tudor protein. Four of the five invariant amino acids (two Asp, one Tyr and one Arg) found in a similar segment of about 80 amino acids in the eight well-defined tudor domains undoubtedly play an essential role in the function of the tudor protein, putatively in an enzymic mechanism involving the polar nature of these residues.

We also report here that the EBNA-2 coactivator p100 consists of a succession of similar domains which probably adopt the SN-fold (SN-like domains, Figure 5) and could share the primary function performed by this enzyme, i.e. nucleic acid binding. This prediction is consistent with the fact that p100 has been shown to bind single-stranded DNA [1]. The SN-fold belongs to the OB-fold superfamily which has been remarkably conserved at the three-dimensional level throughout evolution and is suitable for a large variety of sequences showing little or no apparent relatedness [21]. It includes a wide variety of nucleic acid binding proteins (Table 1) able to recognize diverse forms of nucleic acids (double or single-stranded DNA, tRNA, mRNA) and performing multiple functional roles (transcriptional activation or repression, chromatin modification, DNA repair, translation repression, RNA packaging etc.).

In the light of this function and the known OB-folds, the aromatic and basic residues of the p100 SN-like domains could serve as sites for oligonucleotide binding. The large loop linking strand β_3 and the first α -helix (L3 α 1) is especially rich in arginines and lysines (Figure 4) and therefore seems to be particularly suited to playing an essential role in this function (binding to phosphates). On the other hand, the repetition of the SN-like motif in p100 may also be essential to achieve full biological activity, as suggested by the oligomeric state of several OB-folds (GN5 for example [27]).

SN has been used for many years as a model protein in studies of enzymic catalysis and protein folding. One of the more recent studies has shown that an N-terminal SN fragment corresponding to the OB-fold adopts a stable tertiary structure, provided that two mutations are introduced, probably consolidating the hydrophobic core of the OB-fold subdomain [22]. Interestingly, we report here the existence, in a fifth domain of p100, of a hybrid structure corresponding to the C-terminal SN domain (the two shaded α -helices in Figure 1) following a region lacking the sequence characteristics of the SN OB-fold (subdomain A') but similar to the tudor domain with which it shares the same length. Owing to its similarity with the tudor domain, p100 subdomain A', like the OB-fold, seems to constitute an autonomous folding unit. However, whether this domain corresponds to a highly modified OB-fold, as suggested by its occurrence in a chain of repeated SN-like domains and between two subdomains B, or whether it adopts a different fold, remains to be elucidated.

We acknowledge the ORGANIBIO-CM2AO Structural Biology programme for financial support. Sequence data were analysed using the INFOBIOGEN facilities (<http://kuwww.infobiogen.fr/>) and the Expasy and NCBI servers (<http://expasy.hcuge.ch/> and <http://ncbi.nlm.nih.gov/>).

REFERENCES

- 1 Tong, X., Drapkin, R., Yalamanchili, R., Mosialos, G. and Kieff, E. (1995) Mol. Cell. Biol. **15**, 4735–4744
- 2 Peterson, M. G., Inostroza, J., Maxon, M. E., Flores, O., Admon, A., Reinberg, D. and Tjian, R. (1991) Nature (London) **354**, 369–373

- 3 Maxon, M. E., Goodrich, J. A. and Tjian, R. (1994) *Genes Dev.* **8**, 515–524
- 4 Gaboriaud, C., Bissery, V., Benchetrit, T. and Mornon, J. P. (1987) *FEBS Lett.* **224**, 149–155
- 5 Lemesle-Varloot, L., Henrissat, B., Gaboriaud, C., Bissery, V., Morgat, A. and Mornon, J. P. (1990) *Biochimie* **72**, 555–574
- 6 Woodcock, S., Mornon, J. P. and Henrissat, B. (1992) *Protein Eng.* **5**, 629–635
- 7 Henrissat, B., Callebaut, I., Fabrega, S., Lehn, P., Mornon, J. P. and Davies, G. (1995) *Proc. Natl. Acad. Sci. U.S.A.* **92**, 7090–7094
- 8 Saxena, I. R., Brown, R. M. J., Fevre, M., Geremia, R. A. and Henrissat, B. (1995) *J. Bacteriol.* **177**, 1419–1424
- 9 Thoreau, E., Petridou, B., Kelly, P. A., Djiane, J. and Mornon, J. P. (1991) *FEBS Lett.* **282**, 26–31
- 10 Callebaut, I., Renoir, J. M., Lebeau, M. C., Massol, N., Burny, A., Baulieu, E. E. and Mornon, J. P. (1992) *Proc. Natl. Acad. Sci. U.S.A.* **89**, 6270–6274
- 11 Callebaut, I. and Mornon, J. P. (1995) *FEBS Lett.* **374**, 211–215
- 12 Cotton, F. A., Hazen, E. E. and Legg, M. J. (1979) *Proc. Natl. Acad. Sci. U.S.A.* **76**, 2551–2555
- 13 Loll, P. J. and Lattmann, E. E. (1989) *Proteins: Struct. Funct. Genet.* **5**, 183–201
- 14 Hynes, T. R. and Fox, R. O. (1991) *Proteins* **10**, 92–105
- 15 Chesneau, T. R. and El Solh, N. (1994) *Gene* **145**, 41–47
- 16 Close, S. M. and Kado, C. I. (1992) *Mol. Microbiol.* **6**, 521–527
- 17 Gerlitz, M., Hrabak, O. and Schwab, H. (1990) *J. Bacteriol.* **172**, 6194–6203
- 18 Yoshioka, Y., Fujita, Y. and Ohtsubo, E. (1990) *J. Mol. Biol.* **214**, 39–53
- 19 Bairoch, A., Bucher, P. and Hoffmann, K. (1995) *Nucleic Acids Res.* **24**, 189–196
- 20 Kraulis, P. (1991) *J. Appl. Crystallogr.* **24**, 946–950
- 21 Murzin, A. G. (1993) *EMBO J.* **12**, 861–867
- 22 Alexandrescu, A. T., Gittis, A. G., Abeygunawardana, C. and Shortle, D. (1995) *J. Mol. Biol.* **250**, 134–143
- 23 Wolffe, A. P. (1994) *BioEssays* **16**, 245–251
- 24 Cavarelli, J., Rees, B., Ruff, M., Thierry, J. and Moras, D. (1993) *Nature (London)* **362**, 181–184
- 25 Onesti, S., Miller, A. D. and Brick, P. (1995) *Structure* **3**, 163–176
- 26 Schindelin, H., Jiang, W., Inouye, M. and Heinemann, U. (1994) *Proc. Natl. Acad. Sci. U.S.A.* **91**, 5119–5123
- 27 Skinner, M. M., Zhang, H., Leschnitzer, D. H., Guan, Y., Bellamy, H., Sweet, R. M., Gray, C. W., Konings, R. N. H., Wang, A. H.-J. and Terwilliger, T. C. (1994) *Proc. Natl. Acad. Sci. U.S.A.* **91**, 2071–2075
- 28 Golden, B. L., Hoffman, D. W., Ramakrishnan, V. and White, S. W. (1993) *Biochemistry* **32**, 12812–12820
- 29 Sixma, T. K., Pronk, S. E., Kalk, K. H., Wartna, E. S., van Zanten, B. A. M., Witholt, B. and Hol, W. G. J. (1991) *Nature (London)* **351**, 371–377
- 30 Verlinde, C. L. M. J., Merritt, E. A., van den Akker, F., Kim, H., Feil, I., Delboni, L. F., Mande, S. C., Sarfaty, S., Petra, P. H. and Hol, W. G. J. (1994) *Protein Sci.* **3**, 1670–1686
- 31 Stein, P. E., Boodhoo, A., Tyrrell, G. J., Brunton, J. L. and Read, R. J. (1992) *Nature (London)* **355**, 748–750
- 32 Swaminathan, S., Furey, W., Pletcher, J. and Sax, M. (1992) *Nature (London)* **359**, 801–806
- 33 Parsad, G. S., Earhart, C. A., Murray, D. L., Novick, R. P., Schlievert, P. M. and Ohlendorf, D. H. (1993) *Biochemistry* **32**, 13761–13766
- 34 Williamson, R. A., Martorell, G., Carr, M. D., Murphy, G., Docherty, A. J. P., Freedman, R. B. and Feeney, J. (1994) *Biochemistry* **33**, 11745–11759
- 35 Campos-Olivas, R., Bruix, M., Santoro, J., Lacadena, J., Martinez delPozo, A., Gavilanes, J. G. and Rico, M. (1995) *Biochemistry* **34**, 3009–3022
- 36 Altschul, S., Gish, W., Miller, W., Myers, E. and Lipman, D. (1990) *J. Mol. Biol.* **215**, 403–410
- 37 Pearson, W. R. and Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. U.S.A.* **85**, 2444–2448
- 38 Wilson, R., Ainscough, R., Anderson, K., Baynes, C., Berks, M., Bonfield, J., Burton, J., Connell, M., Copsey, T., Cooper, J. et al. (1994) *Nature (London)* **368**, 32–38
- 39 Landes, C., Henaut, A. and Risler, J. L. (1992) *Nucleic Acids Res.* **20**, 3631–3637
- 40 Weber, D. J., Gittis, A. G., Mullen, G. P., Abeygunawardana, C., Lattman, E. E. and Mildvan, A. S. (1992) *Proteins: Struct. Funct. Genet.* **13**, 275–287
- 41 Boswell, R. E. and Mahowald, A. P. (1985) *Cell* **43**, 97–104
- 42 Golumbeski, G. S., Bardsley, A., Tax, F. and Boswell, R. E. (1991) *Genes Dev.* **5**, 2060–2070
- 43 Bardsley, A., McDonald, K. and Boswell, R. E. (1993) *Development* **119**, 207–219
- 44 Orengo, C. A., Jones, D. T. and Thornton, J. M. (1994) *Nature (London)* **372**, 631–634