A variety of DNA-binding and multimeric proteins contain the histone fold motif

Andreas D. Baxevanis, Gina Arents¹, Evangelos N. Moudrianakis¹ and David Landsman*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, Room 8N-805, Bethesda, MD 20894, USA and ¹Department of Biology, The Johns Hopkins University, Baltimore, MD 21218, USA

Received March 30, 1995; Revised and Accepted June 7, 1995

ABSTRACT

The histone fold motif has previously been identified as a structural feature common to all four core histones and is involved in both histone–histone and histone–DNA interactions. Through the use of a novel motif searching method, a group of proteins containing the histone fold motif has been established. The proteins in this group are involved in a wide variety of functions related mostly to DNA metabolism. Most of these proteins engage in protein–protein or protein–DNA interactions, as do their core histone counterparts. Among these, the CCAAT-specific transcription factor CBF and its yeast homologue HAP are two examples of multimeric complexes with different component subunits that contain the histone fold motif. The histone fold proteins are distantly related, with a relatively small degree of absolute sequence similarity. It is proposed that these proteins may share a similar three-dimensional conformation despite the lack of significant sequence similarity.

INTRODUCTION

The core histone proteins (H2A, H2B, H3 and H4) have been the subject of innumerable biochemical and biophysical studies, primarily due to their central role in the compaction of DNA into nucleosomes (1) and their role in the overall organization of eukaryotic chromosomes [cf. (2,3) for reviews]. The core histones are small, basic proteins (11–15 kDa) that contain a high proportion of positively-charged amino acids, primarily lysine and arginine. These residues help the core histones bind tightly to DNA through ionic interactions with the negative charges along the phosphate backbone of the DNA. One of the most striking features regarding the core histone proteins is their remarkable degree of conservation throughout evolution: for example, the sequences for pea and calf thymus H4 differ at only two of 102 positions (4).

X-ray crystallographic studies at 3.1 Å resolution have confirmed an earlier biochemical proposal (5) that the core histone octamer is a tripartite protein assembly composed of a central (H3–H4)₂ tetramer flanked by two H2A–H2B dimers (6). The shape of this assembly resembles a cylindrical wedge, and the folding of all eight protein chains within it is guided by a novel protein motif, the histone fold (7). The histone fold can be thought of as an extended

helix–strand–helix motif, comprised of a short alpha-helix (~10 residues), followed by a loop/beta-strand segment, a long alpha-helix (~27 residues), a second loop/beta-strand segment, and another short alpha-helix (~10 residues). Head-to-tail association of two histone folds forms the fundamental histone heterodimers, the ultimate building blocks of the core histone octamer. The internal symmetries of each dimer, as well as the highly compact arrangement of the heterodimers within the histone octamer, are dictated by the properties of the histone fold. It has been proposed previously that the appearance of the histone fold in all four of the core histone proteins resulted from a parallel duplication of a single primordial gene encoding a product involved in both protein assembly and DNA binding (7).

In recent years, a small number of proteins have been identified as containing histone-like components within their covalent continuum, such as TAF_{II}40 and TAF_{II}60 from *Drosophila* (8) and macro-H2A from rat (9). Given the extent of the overall conservation of the histones and the prevalence of the histone fold in such an 'early' group of proteins, it appeared reasonable to undertake a computational study to determine the possible presence of the histone fold in other proteins and its utilization in other protein assembly schemes. We report here that the histone fold motif has been identified in numerous DNA-binding proteins, including transcription factors and recombinases. Among these are the CCAAT-specific transcription factor CBF and its yeast homologue HAP, multimeric complexes with different component subunits that contain the histone fold motif. We suggest that these proteins may use principles of assembly similar to those observed with the core histones in order to achieve productive protein–protein interactions, protein–DNA interactions, or both.

MATERIALS AND METHODS

Compilation of core histone sequences

Core histone protein sequences were obtained from the Histone Sequence Database maintained at NCBI (10). The database was compiled by searching the SWISSPROT version 30.0 (11), Protein Identification Resource (PIR) version 41.10 (12), GenPept release 85.0 (13) and the Protein Data Bank (PDB, April 1994) databases using both the BLASTP (14) and FASTA (15) algorithms. The histone database files can be accessed through the World Wide Web (<http://www.ncbi.nlm.nih.gov/Baxevani/HISTONES>), are available for anonymous ftp at the National

* To whom correspondence should be addressed

Figure 1. Behavior of the MoST algorithm as a function of cutoff value R . The number of sequences retrieved by searching the SWISSPROT version 30.0 (11), Protein Identification Resource (PIR) version 41.10 (12), GenPept release 85.0 (13) and the Protein Data Bank (PDB, April 1994) databases is plotted against the number of MoST iterations performed. R values for each plot are shown in the inset and represent the ratio of the expected number of false positives found to the predicted number of true positives found during the database searches. R values were set from 0.001 to 0.05, with seven different values of R being used in each case; only selected data are shown in this figure. Results for values of R that produced divergence rather than convergence were not considered. As the original input alignment blocks were non-redundant, the number of sequences found represents both multiple occurrences of the same histone sequence in the databases, whether they be from the same species or from several different species, as well as non-histone proteins which putatively contain the histone fold.

Center for Biotechnology Information (ncbi.nlm.nih.gov/directory/pub/baxeavanis/histones) or can be obtained from the authors. E-mail to: baxeavani@ncbi.nlm.nih.gov or landsman@ncbi.nlm.nih.gov

Motif searches and sequence analysis

Once all of the histone fold sequences were compiled and aligned, the Motif Search Tool (MoST) (16) was used to detect the histone fold motif within other proteins in the databases. In this study, a Bayesian prediction approach was used to calculate the weight matrices for alignment blocks of histones H2A, H2B, H3 and H4 separately, as well as for an overall alignment block comprised of all four core histones. There were 65 sequences in the H2A alignment block, 66 in the H2B block, 80 in the H3 block, 57 in the H4 block, and 268 in the combined block; the data set was non-redundant. Since the histone fold motif in H4 is shorter than those in the other three core histones by one residue, an 'X' was inserted following residue 19 in the H4 sequences when combined into the overall alignment block (see Fig. 3 for numbering). The ratio R (expected number of false positives/predicted number of true positives) was set from 0.001 to 0.05, with seven different values of R used for each data set. While the MoST searches are generally expected to converge (16), certain critical values of R cause the number of false positives retrieved to produce divergence rather than convergence. Here, values of $R > 0.03$ did not yield convergence in the database searches, and results from these searches were not considered. As a final search step, statistically significant sequence matches found in the course of the MoST searches were run through BLASTP (14) and BLAST3 (17) to detect any additional significant sequence matches; no new sequences were found as a result of these searches. Secondary structure predictions were performed on each sequence found using the profile neural network method PHDsec (18,19) through the PredictProtein E-mail server: (PredictProtein@EMBL-Heidelberg.de).

RESULTS AND DISCUSSION

Motif searches

Upon compilation of histone data sets to be used as the basis for motif searching, the Motif Search Tool (MoST) (16) was used to search the major protein databases for the presence of the histone fold in other proteins. In this method, the sequences are aligned to create an alignment block with N sequences of length L , L in this case being the length of the histone fold. Using this alignment block, a protein weight matrix is generated. The searches are then performed by sliding this weight matrix along all of the sequences in the protein databases. A score is computed for every segment of length L by summing up the appropriate weight matrix elements. As statistically significant matches to the alignment block are found, they are in turn added to the alignment block. At the end of a round of database searches, the weight matrix is recalculated and the database searches repeated. This iterative process continues until no new statistically significant segments are found. The process, therefore, capitalizes on new information obtained in the course of the database searches.

For each histone, seven different values of R ranging from 0.001 to 0.05 were used in performing the motif searches, with the iterative behavior of the MoST algorithm varying from search to search (Fig. 1). One would expect these searches to return

Figure 2. Multiple sequence alignment of proteins containing the histone fold motif detected using alignment blocks of individual histones. 65 input sequences were in the H2A alignment block, 66 in the H2B block, 80 in the H3 block and 57 in the H4 block; the data set was non-redundant. The abbreviation for each protein corresponds to those listed in Table 1. For each core histone, the major human histone sequence is shown at the top of each individual alignment. At each position, residues in agreement with those in the first sequence of the individual alignment are shown in inverse type. ALSCRIPT version 2.0 (56) was used to format the alignment.

additional histone sequences that were not in the original non-redundant alignment blocks, as well as sequences for non-histone proteins which putatively contain the histone fold. In the case of the H2A alignment block, which consisted of a non-redundant set of 65 sequences, lower values of R (≤ 0.01) produced fairly similar results, with the searches concluding after two or three iterations. For $R \geq 0.02$, however, the number of false positives retrieved increased to such an extent that the search process became divergent rather than convergent. In each case where divergence was observed, the results from that particular search were discarded. The maximum value of R producing convergence varied from 0.001 for histone H4 to 0.03 for histone H3. While a large number of sequences appear to be found in all cases, the majority of those sequences represent multiple occurrences of the same histone sequence in the databases, whether they be from the same species or from several different species; recall that the initial input set was non-redundant.

An alignment of the histone fold regions for the proteins identified through these individual histone searches is presented in Figure 2. The feature which becomes immediately apparent upon examination of these alignments is that there is no overwhelming similarity between the sequences of the identified proteins and those of the human histone sequences shown in the alignment. In the case of H4, the MoST searches identified only four archaeal DNA binding proteins; while the archaeal proteins are almost identical to one another, there is only absolute agreement between these proteins and human H4 in 13 out of 64 positions. For H3, again only four sequences were identified, all of which are H3-like fusion proteins; here, seven of the 65 positions are identical to the

sequence of human H3. In the case of H2A, only two positions agree amongst all found sequences (³⁵L and ³⁶E). For H2B, there is no absolutely conserved position in the alignment. Across all four of these alignments, only two positions at the N-terminal end and five positions at the C-terminal end are particularly lysine- or arginine-rich.

MoST searches were also performed on a combined data set consisting of all four core histones aligned in the histone fold region (Fig. 1, bottom panel). The results of these searches were similar to those performed with the individual histone alignment blocks, with this non-redundant alignment block of 268 sequences producing convergence within five iterations for $R \leq 0.03$. However, the discriminatory power of this combined alignment block was much greater than that of the individual histones: the combined block identified 32 sequences as containing the histone fold motif, more than were identified by any one of the individual histone alignment blocks (Table 1). An alignment of the histone fold regions for the proteins found using this combined data set is presented in Figure 3. As with the results of the individual histone alignment block searches described above, few positions are seen to be particularly lysine- or arginine-rich.

Given that the sequences of this obviously distantly-related family of proteins are not highly conserved, consideration must be given to the three-dimensional structure of these proteins. It is generally accepted that structure is conserved to a greater extent than is sequence and that there is a limited number of backbone motifs (20,21). Therefore, similarities between proteins may not necessarily be obvious by examining sequence alignments. An example of this is the relationship between the SIN1/SPT2

Figure 3. Multiple sequence alignment of proteins containing the histone fold motif detected using an alignment block consisting of all four core histones. 268 input sequences were in the combined alignment block; the data set was non-redundant. The abbreviation for each protein corresponds to those listed in Table 1. The major sequence for all four human core histones are shown at the top of the alignment. At each position, residues in agreement with any one of the four core histones are shown in inverse type. ALSCRIPT version 2.0 (56) was used to format the alignment.

protein in *Saccharomyces cerevisiae* (22) and the HMG-1 box motif. Here, only 10% of the residues of SIN1/SPT2 are identical to those of rat HMG-1 box 2 but there is a statistically significant fit of sequence to structure as assessed by homology model building (23). A similar theme may be occurring within the family of histone fold proteins.

Histone fold motif search results

The 40 proteins identified in this study as possibly containing the histone fold motif are involved in a variety of cellular functions. Most of the proteins in this family are involved in processes such as transcriptional regulation or DNA compaction. The higher-order interactions in which each of these proteins are involved are summarized in Table 1. Based on these interactions, the proteins in this group can be broken into three general classes, as described below.

Protein-protein and protein-DNA/RNA interactions. All but two of the putative histone fold proteins fall into this class, as would be expected based on the structure and function of the histone proteins themselves. The most closely-related of these proteins to the histones are the archaeal DNA-binding proteins HMf, HMt and HMv (24–26). These proteins bind to double-stranded DNA and increase the resistance of bound DNA to thermal denaturation, an important feature since the optimal growth temperature for *Methanothermobacter fervidus* is 83°C (25). DNA binding is through a heterodimeric complex of two slightly different polypeptides (e.g. HMf-1 and HMf-2), yielding what has been termed ‘nucleosome-like structures’. Based on biochemical data, a model for DNA compaction by the HMf proteins has been proposed (27): At low protein-to-DNA mass ratios, the binding of the HMf heterodimer to DNA causes a local unwinding of the DNA. At higher mass ratios, interactions between bound HMf

heterodimers occur, leading to the formation of nucleosome-like structures which constrain 90–150 bp of DNA in about 1.5 positive supercoils. The histones, in contrast, lead to the formation of negative supercoils.

As previously stated, a large number of proteins identified as containing the histone fold are involved in transcriptional regulation. As is the case with the histones, these proteins both associate with DNA and are assembled in some sort of protein complex. Amongst the most interesting of these are CBF-A and CBF-C which, together with CBF-B, form the CCAAT-specific transcription factor CBF (or NF-Y). CBF binds to promoters controlling $\alpha 1(I)$ and $\alpha 2(I)$ collagen, albumin, MHC class II products and β -actin, among others (28). Recent work on CBF indicates that CBF-A and CBF-C form a complex to which CBF-B can then bind, CBF-B being unable to bind to either CBF-A or CBF-C separately (29). Binding to DNA is then primarily through interactions with CBF-B, though additional interactions are thought to occur through both CBF-A and CBF-C (29,30). In the course of the MoST searches, CBF-C was detected using the H2B alignment block, while CBF-A was detected using the combined alignment block. It is possible, then, that these two subunits are utilizing slightly different forms of the histone fold motif in forming the ternary complex, a situation similar to what would be expected within the core histone octamer itself. A parallel case exists in yeast, where the HAP2–HAP3–HAP5 complex binds to the CCAAT-box region of one of the upstream activation sites of the *CYC1* gene; this gene encodes for iso-1-cytochrome *c* and is implicated in the global activation of cytochrome and related genes (30–32). Here, HAP3 and HAP5 first form a dimer to which HAP2 can then bind; this trimer then binds to DNA primarily through a DNA-binding domain on HAP2 (30). The evolutionary relatedness of CBF and HAP are shown by the ability of CBF-C to substitute for HAP5 within the yeast complex (29). As with the CBF proteins, HAP5 was detected using the H2B alignment block,

Table 1. Histone fold motif search results

Code	Description	Organism	Accession Number	MoST Searches ¹					Known Interactions ²			References
				H2A	H2B	H3	H4	Comb.	Pro-Pro	Pro-NA	DNA Met.	
chk-CBF-A	Transcription factor CBF-A	<i>Gallus gallus</i>	splP25207					•	•	•	•	–
hum-CBF-A	Transcription factor CBF-A	<i>Homo sapiens</i>	splP25208					•	•	•	•	(57, 58)
mus-CBF-A	Transcription factor CBF-A	<i>Mus musculus</i>	splP22569					•	•	•	•	(57-59)
rat-CBF-A	Transcription factor CBF-A	<i>Rattus norvegicus</i>	pirA35777					•	•	•	•	(28)
pma-CBF-A	Transcription factor CBF-A	<i>Petromyzon marinus</i>	splP25210					•	•	•	•	(58)
rat-CBF-C	Transcription factor CBF-C	<i>Rattus norvegicus</i>	gplU17607		•				•	•	•	(29)
pce-DGDR	Insect-type dehydrogenase	<i>Pseudomonas cepacia</i>	gplJ05282		•				•			(60)
sso-DNApol	DNA-directed DNA polymerase	<i>Sulfolobus solfataricus</i>	pirS34120						•	?	•	(61)
hum-DR1	TATA-binding protein-associated phosphoprotein	<i>Homo sapiens</i>	splQ01658		•				•		•	(39)
kma-HAP3	Transcription activator HAP3	<i>Kluyveromyces marxianus</i>	pirS41926					•	•	•	•	–
sce-HAP3	Transcription activator HAP3	<i>Saccharomyces cerevisiae</i>	splP13434					•	•	•	•	(30, 31, 62)
sce-HAP5	Transcription activator HAP5	<i>Saccharomyces cerevisiae</i>	gplU19932		•				•	•	•	(63)
mfe-HMf1	DNA-binding protein HMf-1	<i>Methanothermobacter fervidus</i>	pirA47036	•	•		•	•	•	•	•	(24)
mfe-HMf2	DNA-binding protein HMf-2	<i>Methanothermobacter fervidus</i>	pirA35959	•	•		•	•	•	•	•	(25)
mth-HMt2	DNA-binding protein HMt-2	<i>Methanobacterium thermoautotrophicum</i>	pirC47036	•	•		•	•	•	•	•	(24)
mvo-HMvA	DNA-binding protein HMv-A	<i>Methanococcus voltae</i>	splQ03576		•			•	•	•	•	(26)
mfo-JF1	Archaeal histone JF-1	<i>Methanobacterium formicicum</i>	gplU12930	•	•		•	•	•	•	•	–
mfo-J0316	Hypothetical protein J0316	<i>Saccharomyces cerevisiae</i>	gplX77688					•				(64)
eco-RcsB	Capsid synthesis regulator B-component	<i>Escherichia coli</i>	splP14374					•	•	?	•	(36)
eco-Rec	Recombinase	<i>Escherichia coli</i>	splP13606					•	•	?	•	(65)
ppu-Rec	Recombinase	<i>Pseudomonas putida</i>	splP30739					•	•	?	•	(66)
avi-rpoN	Sigma (rpoN) modulation protein	<i>Azotobacter vinelandii</i>	splP17160		•				•	•	•	(67, 68)
pae-rpoN	Sigma (rpoN) modulation protein	<i>Pseudomonas aeruginosa</i>	gplL26916		•				•	•	•	(37)
ppu-rpoN	Sigma (rpoN) modulation protein	<i>Pseudomonas putida</i>	splP15592		•				•	•	•	(69, 70)
ppu-rpoX	rpoX protein	<i>Pseudomonas putida</i>	pirS06713		•				•	•	•	–
dme-SOS	Son of sevenless	<i>Drosophila melanogaster</i>	splP26675	•	•			•	•	•	•	(38, 71)
hum-SOS-1	Son of sevenless 1	<i>Homo sapiens</i>	pirA37488	•	•			•	•	•	•	(72)
mus-SOS-1	Son of sevenless 1	<i>Mus musculus</i>	gplZ11574	•					•	•	•	(73)
mus-SOS-2	Son of sevenless 2	<i>Mus musculus</i>	pirS25714	•	•				•	•	•	(73)
dme-TAFII40	TAFII40 gene product	<i>Drosophila melanogaster</i>	gplL29540					•	•	•	•	(8, 74)
dme-TAFII60	TAFII60 gene product	<i>Drosophila melanogaster</i>	gplL25443					•	•	•	•	(8, 33)
hum-TFIIB	Transcription factor IIB	<i>Homo sapiens</i>	splQ00403					•	•	•	•	(34, 75)
rat-TFIIB	Transcription factor IIB	<i>Rattus norvegicus</i>	splP29053					•	•	•	•	(76)
xla-TFIIB	Transcription factor IIB	<i>Xenopus laevis</i>	splP29054					•	•	•	•	(77)
eco-TNPR	Transposon TN3 resolvase	<i>Escherichia coli</i>	splP03011					•	•	?	•	(78, 79)
kpn-TNPR	Transposon TN1331 resolvase	<i>Klebsiella pneumoniae</i>	splP21424					•	•	?	•	(80, 81)
sce-YKE9	H3-like protein	<i>Saccharomyces cerevisiae</i>	splP36012			•		•		?		(82)
cel-YL82	H3-like protein in chromosome III	<i>Caenorhabditis elegans</i>	splP34440		•	•		•		?		(83)
cel-YMH3	H3-like protein in chromosome III	<i>Caenorhabditis elegans</i>	splP34470			•		•		?		(83)
lam-Y194	Hypothetical protein ORF194	Bacteriophage lambda	splP03740			•						(84)
Total Matches				8	17	4	4	32				

¹Bullets indicate which MoST searches produced a statistically significant hit of the alignment block to the databases. The columns H2A, H2B, H3 and H4 represent the searches done using alignment blocks composed of individual histones, while the Comb. column represents the database searches done with the combined alignment block comprised of all four core histones.

²Bullets indicate proteins with known protein-protein or protein-DNA/RNA interactions, as well as proteins involved in DNA metabolism. Question marks indicate possible or probable interactions that have not been directly demonstrated.

while HAP3 was detected using the combined alignment block. While both CBF-B and HAP2 are involved in the formation of these trimeric complexes, the histone fold motif was not detected in either of these proteins by our methods.

The TAFII40 and TAFII60 subunits of the TFIID transcriptional complex also contain the histone fold. TFIID itself contains the

TATA-binding protein (TBP) and a number of TBP-associated factors, or TAFs. TAFII40 has been shown to bind to TAFII60, which may in turn promote the association of other TAFs (33). It has been proposed that the TAFII40-TAFII60 may act much like (H3-H4)₂ in the formation of nucleosome-like structures, as well as in maintaining a semi-compacted state near the start site of

transcription (8). TFIIB, which associates with a complex of TFIID and TFIIA to form an intermediate complex which interacts with RNA polymerase II, also belongs to this group. TFIIB has been suggested to contain two functional domains, one which binds to both DNA and transcription factors, the other which binds to activators (34). A DNA-directed DNA polymerase (DNApol) from the archaeobacterium *Sulfolobus* was also identified; this protein possesses conserved sequence motifs found in a number of family A and B DNA polymerases which are thought to be involved in the 3'→5' exonuclease activity of these enzymes (35).

The positive regulator of colanic acid capsular polysaccharide synthesis, RcsB, is thought to act through protein-protein interactions in stimulating the expression of the *cps* genes coding for these capsule proteins (36). A short DNA-binding helix-turn-helix motif in RcsB (20 residues) has also been observed (36). Another regulatory gene, *rpoN*, encodes a product named σ^{54} , which in turn controls the transcription of the gene encoding for the pilin subunit. The pili are of biological importance, as they increase the virulence of Gram-negative bacteria such as *Pseudomonas* by mediating adhesion to epithelial cells (37). The product of the *son of sevenless* gene, SOS, is a guanine nucleotide exchange factor for Ras1: SOS binds inactive Ras1, forming an intermediate SOS-Ras1 complex in which GDP is exchanged for GTP and the SOS protein is released (38). Rec and TNPR both belong to the 'resolvase' family of site-specific recombinases and are involved in replicative transposition. In this process, a transposable element is duplicated, with the duplicate copy being used in the recombination event, the final recombination being mediated by the resolvase. Rec and TNPR are highly similar and both contain a 20-residue helix-turn-helix region. It is also thought that the serine at position 10 (numbering of whole protein) can transiently form a covalent linkage to DNA during strand cleavage and rejoining.

Protein-protein interactions, with involvement in DNA metabolism. The TATA-binding protein-associated phosphoprotein DR1 is the only protein in the list of identified proteins which is involved in DNA metabolism without binding to DNA. DR1 interferes with the association of TFIIA and/or TFIIB with TBP, thereby inhibiting the formation of a transcriptionally competent complex. Far-Western analysis has shown that the phosphorylation of DR1 regulates its interaction with TBP directly and in the absence of DNA (39).

Protein-protein interactions, with no involvement in DNA metabolism. One protein which possibly contains the histone fold is involved in protein-protein interactions like the histones, but is not thought to bind to DNA and is not involved in DNA regulation or metabolism. DGDR is a pyridoxal 5'-phosphate-dependent enzyme catalyzing both decarboxylation and amino transfer in soil bacteria; it is composed of four identical 47 kDa subunits and is thought to break down cytotoxic peptides produced by soil fungi (40,41).

Proteins unlikely to contain the histone fold motif. In addition to the proteins discussed above, a number of additional sequences were retrieved by the MoST algorithm. However, based on their predicted secondary structure (18,19), these proteins appear unlikely to contain the histone fold motif. Given that the crystal structure of the core histone octamer indicates that the histone fold is highly helical in nature, the application of this second set of criteria provides a physical basis by which possible false positives

can be identified. Coincidentally, none of these proteins have been shown to interact with DNA, and all but three have not been demonstrated to participate in any sort of protein-protein interactions. These putative false positives include the L15 ribosomal subunit from *Chlamidia* (42), the nitrate reductase-linked formate dehydrogenase (43-45), the phycocyanin rod linker protein PYR1 (46), the multi-drug resistance P-glycoprotein (gpL29484), the cell wall-degrading enzyme polygalactouronase (47), the RNA2 polyprotein from grapevine fanleaf virus (48), the bacteriophage resistance protein HAU3 from *Streptomyces* (pir[S42059]), the KA6L gene product from human herpesvirus 6 (gpL16947) and protein RF5 from Herpes simplex virus type 6 (49).

One last group of proteins identified through the MoST searches are transmembrane proteins: the water stress-induced tonoplast intrinsic protein WSI-TIP, which is involved in small molecule transport during osmoregulation and the vacuolation of expanding cells (50), and the turgor-responsive proteins PMIP-1a, -2a and -2b, TMP-A and TMP-C from corn and *Arabidopsis* (51,52). While the secondary structure predictions (18,19) indeed indicate that these proteins have the potential to form large helical regions (as would be expected for any transmembrane protein), their cellular localization most likely makes these false positives.

Future perspectives

Through the use of a novel motif searching method, a group of proteins putatively containing the histone fold has been established. Most of these proteins are engaged in both protein-protein and protein-DNA interactions, as are the histones themselves. As has been speculated with TAF_{II}40 and TAF_{II}60 (8), the proteins with known nucleic acid and subunit interactions may be involved in the formation of variant nucleosome-like structures or in keeping distant regions of DNA together during processes such as transcription.

The existence of a common folding motif in a variety of proteins of significantly different function leads to the question of how this motif is being used in these alternative contexts. Alpha-enolase and tau-crystallin provide an excellent example of how stable protein structures can be used in alternative biological settings: These proteins, which are identical and encoded by a single gene, function as both a glycolytic enzyme in the liver as well as a structural protein within the lens of the eye (53,54). The application of computational methods such as homology model building (55), which has already been applied to protein families such as the high mobility group (23), has the potential to supply valuable structural information regarding the nature of the histone fold in the absence of three-dimensional information for each of these proteins. Results of such studies will be of great interest in better establishing the heretofore obscure relationships between proteins discovered to have a common folding domain.

ACKNOWLEDGEMENTS

We wish to thank Drs John Wootton and Eugene Koonin for their critical review of this manuscript.

REFERENCES

- 1 Kornberg, R. and Thomas, J.O. (1974) *Science*, **184**, 865-868.
- 2 van Holde, K.E. (1989) *Chromatin*. Springer-Verlag, New York.
- 3 Wolffe, A. (1992) *Chromatin: Structure and Function*. Academic Press, San Diego.

- 4 DeLange, R.J., Fambrough, D.M., Smith, E.L. and Bonner, J. (1969) *J. Biol. Chem.*, **244**, 5669–5679.
- 5 Eickbush, T.H. and Moudrianakis, E.N. (1978) *Biochemistry*, **17**, 4955–4964.
- 6 Arents, G., Burlingame, R.W., Wang, B.-C., Love, W.E. and Moudrianakis, E.N. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 10148–10152.
- 7 Arents, G. and Moudrianakis, E.N. (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 10489–10493.
- 8 Kokubo, T., Gong, D.W., Wootton, J.C., Horikoshi, M., Roeder, R.G. and Nakatani, Y. (1994) *Nature*, **367**, 484–487.
- 9 Pehrson, J.R. and Fried, V.A. (1992) *Science*, **257**, 1398–1400.
- 10 Baxevanis, A.D. and Landsman, D. (1996) *Nucleic Acids Res.*, **24**, in press.
- 11 Bairoch, A. and Boeckmann, B. (1993) *Nucleic Acids Res.*, **21**, 3093–3096.
- 12 Barker, W.C., George, G.D., Mewes, H.-W., Pfeiffer, F. and Tsugita, A. (1993) *Nucleic Acids Res.*, **21**, 3089–3092.
- 13 Benson, D., Lipman, D.J. and Ostell, J. (1993) *Nucleic Acids Res.*, **21**, 2963–2965.
- 14 Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
- 15 Pearson, W.R. and Lipman, D.J. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 2444–2448.
- 16 Tatusov, R.L., Altschul, S.F. and Koonin, E.V. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 12091–12095.
- 17 Karlin, S. and Altschul, S.F. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 2264–2268.
- 18 Rost, B. and Sander, C. (1993) *J. Mol. Biol.*, **232**, 584–599.
- 19 Rost, B. and Sander, C. (1994) *Proteins*, **19**, 55–72.
- 20 Chothia, C. and Lesk, A.M. (1986) *EMBO J.*, **5**, 823–826.
- 21 Chothia, C. (1992) *Nature*, **357**, 543–544.
- 22 Kruger, W. and Herskowitz, I. (1991) *Mol. Cell. Biol.*, **11**, 4135–4146.
- 23 Baxevanis, A.D., Bryant, S.H. and Landsman, D. (1995) *Nucleic Acids Res.*, **23**, 1019–1029.
- 24 Tabassum, R., Sandman, K.M. and Reeve, J.N. (1992) *J. Bacteriol.*, **174**, 7890–7895.
- 25 Sandman, K., Krzycki, J.A., Dobrinski, B., Lurz, R. and Reeve, J.N. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 5788–5791.
- 26 Agha-Amiri, K. and Klein, A. (1993) *Nucleic Acids Res.*, **21**, 1491.
- 27 Grayling, R.A., Sandman, K. and Reeve, J.N. (1994) *System. Appl. Microbiol.*, **16**, 582–590.
- 28 Vuorio, T., Maity, S.N. and de Crombrughe, B. (1990) *J. Biol. Chem.*, **265**, 22480–22486.
- 29 Sinha, S., Maity, S.N., Lu, J. and de Crombrughe, B. (1995) *Proc. Natl. Acad. Sci. USA*, **92**, 1624–1628.
- 30 Xing, Y., Fikes, J.D. and Guarente, L. (1993) *EMBO J.*, **12**, 4647–4655.
- 31 Hahn, S., Pinkham, J., Wei, R., Miller, R. and Guarente, L. (1988) *Mol. Cell. Biol.*, **8**, 655–663.
- 32 Olesen, J., Hahn, S. and Guarente, L. (1993) *Cell*, **51**, 953–961.
- 33 Weinzierl, R.O., Ruppert, S., Dynlacht, B.D., Tanese, N. and Tjian, R. (1993) *EMBO J.*, **12**, 5303–5309.
- 34 Ha, I., Lane, W.S. and Reinberg, D. (1991) *Nature*, **352**, 689–695.
- 35 Pisani, F.M., De Martino, C. and Rossi, M. (1992) *Nucleic Acids Res.*, **20**, 2711–2716.
- 36 Stout, V. and Gottesman, S. (1990) *J. Bacteriol.*, **172**, 659–669.
- 37 Jin, S., Ishimoto, K. and Lory, S. (1994) *J. Bacteriol.*, **176**, 1316–1322.
- 38 Simon, M.A., Botwell, D.D.L., Dodson, G.S., Laverty, T.R. and Rubin, G.M. (1991) *Cell*, **67**, 701–716.
- 39 Inostroza, J.A., Mermelstein, F.H., Ha, I., Lane, W.S. and Reinberg, D. (1992) *Cell*, **70**, 477–489.
- 40 Brückner, H. and Przybylski, M. (1984) *J. Chromatog.*, **296**, 263–275.
- 41 Schmitt, H. and Jung, G. (1985) *Liebigs Ann. Chem.*, **1985**, 321–344.
- 42 Kaul, R., Gray, G.J., Koencke, N.R. and Gu, L.J. (1992) *J. Bacteriol.*, **174**, 1205–1212.
- 43 Zinoni, F., Birkmann, A., Stadtman, T.C. and Bock, A. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 4650–4654.
- 44 Blattner, F.R., Burland, V.D., Plunkett, G., Sofia, H.J. and Daniels, D.L. (1993) *Nucleic Acids Res.*, **21**, 5408–5417.
- 45 Axley, M.J., Grahame, D.A. and Stadtman, T.C. (1990) *J. Biol. Chem.*, **265**, 18213–18218.
- 46 De Lorimier, R., Guglielmi, G., Bryant, D.A. and Stevens, S.E. (1990) *Arch. Microbiol.*, **153**, 541–549.
- 47 Lee, E., Speirs, J., Gray, J. and Brady, C.J. (1990) *Plant. Cell. Env.*, **13**, 513–521.
- 48 Serghini, M.A., Fuchs, M., Pinck, M., Reinbolt, J., Walter, B. and Pinck, L. (1990) *J. Gen. Virol.*, **71**, 1433–1441.
- 49 Josephs, S.F., Ablashi, D.V., Salahuddin, S.Z., Jagodzinski, L., Wong-Staal, F. and Gallo, R.C. (1991) *J. Virol.*, **65**, 5597–5604.
- 50 Ludevid, D., Hoefte, H., Himmelblau, E. and Chrispeels, M.J. (1992) *Plant Physiol.*, **100**, 1633–1639.
- 51 Shagan, T. and Bar-Zvi, D. (1993) *Plant Physiol.*, **101**, 1397–1398.
- 52 Kinoshita, T., Hara-Nishimura, I., Shiraishi, H., Okada, K., Shimura, Y. and Nishimura, M. (1994) *Plant Physiol.*, **105**, 1441–1442.
- 53 Wistow, G. and Piatigorsky, J. (1987) *Science*, **236**, 1554–1556.
- 54 Wistow, G.J., Lietman, T., Williams, L.A., Stapel, S.O., de Jong, W.W., Horwitz, J., et al. (1988) *J. Cell Biol.*, **107**, 2729–2736.
- 55 Bryant, S.H. and Lawrence, C.E. (1993) *Proteins*, **16**, 92–113.
- 56 Barton, G.J. (1993) *Protein Engng*, **6**, 37–40.
- 57 Li, X.-Y., van Huijsduijnen, R., Mantovani, R., Benoist, C.O. and Mathis, D. (1992) *J. Biol. Chem.*, **267**, 8984–8990.
- 58 Li, X.-Y., Mantovani, R., van Huijsduijnen, R., Andre, I., Benoist, C. and Mathis, D. (1992) *Nucleic Acids Res.*, **20**, 1087–1091.
- 59 van Huijsduijnen, R., Li, X.-Y., Black, D., Matthes, H., Benoist, C. and Mathis, D. (1990) *EMBO J.*, **9**, 3119–3127.
- 60 Keller, J.W., Baurick, K.B., Rutt, G.C. and O'Malley, M.V. (1990) *J. Biol. Chem.*, **265**, 5531–5539.
- 61 Prangishvili, D. and Klenk, H.-P. (1993) *Nucleic Acids Res.*, **21**, 2768.
- 62 van Dyck, L., Pearce, D.A. and Sherman, F. (1994) *J. Biol. Chem.*, **269**, 238–242.
- 63 McNabb, D.S., Xing, Y. and Guarente, L. (1995) *Genes Dev.*, **9**, 47–58.
- 64 Purnelle, B., Coster, F. and Goffeau, A. (1994) *Yeast*, **10**, 1235–1249.
- 65 Dodd, H.M. and Bennett, P.M. (1987) *J. Gen. Microbiol.*, **133**, 2031–2039.
- 66 Assinder, S.J., de Marco, P., Sayers, J.R., Shaw, L.E., Winson, M.K. and Williams, P.A. (1992) *Nucleic Acids Res.*, **20**, 5476.
- 67 Merrick, M.J., Gibbins, J. and Toukdarian, A. (1987) *Mol. Gen. Genet.*, **210**, 323–330.
- 68 Merrick, M.J. and Coppard, J.R. (1989) *Mol. Microbiol.*, **3**, 1765–1775.
- 69 Koehler, T., Cayrol, J.M., Ramos, J.L. and Harayama, S. (1989) *Nucleic Acids Res.*, **17**, 10125.
- 70 Inouye, S., Yamada, M., Nakazawa, A. and Nakazawa, T. (1989) *Gene*, **85**, 145–152.
- 71 Bonfini, L., Karlovich, C.A., Dasgupta, C. and Banerjee, U. (1992) *Science*, **255**, 603–606.
- 72 Chardin, P., Camonis, J.H., Gale, N.W., van Aelst, L., Schlessinger, J., Wigler, M.H., et al. (1993) *Science*, **260**, 1338–1343.
- 73 Botwell, D., Fu, P., Simon, M. and Senior, P. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 6511–6515.
- 74 Goodrich, J.A., Hoey, T., Thut, C.J., Admon, A. and Tjian, R. (1993) *Cell*, **75**, 519–530.
- 75 Malik, S., Hisatake, K., Sumimoto, H., Horikoshi, M. and Roeder, R.G. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 9553–9557.
- 76 Tsuboi, A., Conger, K., Garrett, K.P., Conaway, R.C., Conaway, J.W. and Arai, N. (1992) *Nucleic Acids Res.*, **20**, 3250.
- 77 Hisatake, K., Malik, S., Roeder, R.G. and Horikoshi, M. (1991) *Nucleic Acids Res.*, **19**, 6639.
- 78 Heffron, F., McCarthy, B.J., Ohtsubo, H. and Ohtsubo, E. (1992) *Cell*, **18**, 1153–1163.
- 79 Mabilat, C., Lourenco-Vital, J., Goussard, S. and Courvalin, P. (1992) *Mol. Gen. Genet.*, **235**, 113–121.
- 80 Tolmasky, M.E. (1990) *Plasmid*, **24**, 218–226.
- 81 Nobuta, K., Tolmasky, M.E., Crosa, L.M. and Crosa, J.H. (1988) *J. Bacteriol.*, **170**, 3769–3773.
- 82 Purnelle, B., Tettelin, H., van Dyck, L., Skala, J. and Goffeau, A. (1993) *Yeast*, **9**, 1379–1384.
- 83 Wilson, R., Ainscough, R., Anderson, K., Baynes, C., Berks, M., Bonfield, J., et al. (1994) *Nature*, **368**, 32–38.
- 84 Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F. and Petersen, G.B. (1982) *J. Mol. Biol.*, **162**, 729–773.