

# Prediction of an HMG-Box Fold in the C-Terminal Domain of Histone H1: Insights Into Its Role in DNA Condensation

M.M. Srinivas Bharath,<sup>1</sup> Nagasuma R. Chandra,<sup>2</sup> and M.R.S. Rao<sup>1\*</sup>

<sup>1</sup>Department of Biochemistry, Indian Institute of Science, Bangalore, India

<sup>2</sup>Bioinformatics Centre, Indian Institute of Science, Bangalore, India

**ABSTRACT** In eukaryotes, histone H1 promotes the organization of polynucleosome filaments into chromatin fibers, thus contributing to the formation of an important structural framework responsible for various DNA transaction processes. The H1 protein consists of a short N-terminal “nose,” a central globular domain, and a highly basic C-terminal domain. Structure prediction of the C-terminal domain using fold recognition methods reveals the presence of an HMG-box-like fold. We recently showed by extensive site-directed and deletion mutagenesis studies that a 34 amino acid segment encompassing the three S/TPKK motifs, within the C-terminal domain, is responsible for DNA condensing properties of H1. The position of these motifs in the predicted structure corresponds exactly to the DNA-binding segments of HMG-box-containing proteins such as Lef-1 and SRY. Previous analyses have suggested that histone H1 is likely to bend DNA bound to the C-terminal domain, directing the path of linker DNA in chromatin. Prediction of the structure of this domain provides a framework for understanding the higher order of chromatin organization. *Proteins* 2002;49:71–81.

© 2002 Wiley-Liss, Inc.

**Key words:** HMG-box fold; histone H1; DNA condensation; C-terminal domain; chromatin organization

## INTRODUCTION

The larger role of chromatin in the regulation of DNA transaction processes, such as gene expression and replication, apart from its DNA packaging function is now well recognized. The DNA in an eukaryotic cell nucleus exists as beaded filaments with a diameter of 10 nm at low salt concentrations but exist as thicker highly condensed structures popularly known as 30-nm fibers, at physiological salt concentrations.<sup>1</sup> The four core histones wrap DNA around them to generate the fundamental unit of chromatin, the nucleosome core particle.

A continuing central question in chromatin research is to elucidate how the fibers fold and unfold to allow DNA transaction processes such as transcription and replication. The crystal structure of the nucleosome, solved a few years ago, has shown how the core histones form an octamer encapsulated by 146 bp of DNA in nearly two gyres of duplexes.<sup>2</sup> It is known that the 30-nm fibers

require the participation of not only the four core histones H2A, H2B, H3, and H4 but also the linker histone H1 for their formation and/or stabilization.<sup>3,4</sup> The exact role of the linker histone H1 in this process, however, still remains poorly understood. Knowledge of the structure of H1 is essential for probing the molecular mechanisms of the expected DNA condensation. The histone H1 protein consists of three distinct domains, a small 34-residue N-terminal fragment (nose), the central 74-residue globular domain (head), and a slightly larger 110 residue C-terminal domain (tail). In aqueous media, at physiological pH and ionic strength, the N-terminal nose is believed to have no regular structure, whereas the C-terminal domain, which is otherwise unfolded, folds into an ordered structure in the presence of trifluoroethanol.<sup>5</sup> The structure of the globular domain has been well characterized both by NMR<sup>6</sup> and X-ray crystallography,<sup>7</sup> whereas virtually nothing is known about the structure of the C-terminal domain. Biochemical studies and electron microscopic studies have indicated strongly that the C-terminal domain is crucial for DNA condensation<sup>8</sup> and chromatin folding<sup>4,9</sup> to generate a 30-nm chromatin fiber.

Recent advances in bioinformatics and computational biology have made sequence analysis and sequence-based structure predictions a tangible approach for many proteins, not only where significant similarities to a known structure exist but also where no structural templates are obvious from sequence similarities alone.<sup>10</sup> This is apparent from the CASP4 experiment,<sup>11</sup> the most recent of a series of communitywide structure-prediction analyses that show several correct predictions for the target proteins in the test. The predictions can be particularly effective when combined with the knowledge from biochemical and biophysical experiments for that family of proteins. Recent literature shows that such predictions have provided significant insights into function of those protein molecules. For example, prediction of the fold of GGDEF domain present in many prokaryotic proteins has led to the identification of their role as regulatory enzymes

Grant sponsor: Council of Scientific and Industrial Research and Department of Biotechnology, New Delhi.

\*Correspondence to: Prof. M.R.S. Rao, Department of Biochemistry, Indian Institute of Science, Bangalore 560012, India. E-mail: mrsrao@biochem.iisc.ernet.in

Received 31 January 2002; Accepted 25 April 2002

involved in nucleotide cyclization.<sup>12</sup> In the absence of experimental structure determination of the C-terminal domain of H1, we have used a combination of bioinformatics methods and the knowledge from the available biochemical and biophysical data to predict the structure of the C-terminal domain. We also report the structure-function insights gained by the prediction.

## MATERIALS AND METHODS

### Overall Sequence Analysis and Secondary Structure Prediction

The sequence of rat histone H1d, P15865 obtained from the SWISSPROT database was used for all analyses. During sequence analyses, the low-complexity filtering was disabled, because as much as 40% of the C-terminal domain was comprised of lysines. Multiple alignments were performed by using CLUSTALW.<sup>13</sup> Apart from the composition bias, the sequence also contained internal repeats of short stretches, leading to the possibility of frame shift errors. In view of this finding, particular care was taken to evaluate alignments at all stages. Therefore, the multiple alignments obtained through CLUSTALW required editing to maximize the overlap of the known functional motifs (S/TPKK) in all known mammalian linker histones. The PRINTS database<sup>14</sup> was used to analyze sequence profiles. In view of the high compositional bias present in the sequence, secondary structure prediction was conducted by using several well-known methods working on different principles and a consensus obtained with use of the Network Protein sequence analysis server available at <http://npsa-pbil.ibcp.fr/>.<sup>15</sup> The different methods used are as follows: the GORIV,<sup>16</sup> a secondary structure prediction method that uses all possible pair frequencies within windows of 17 residues; HNN,<sup>17</sup> a multivariate linear regression method embedded in a hierarchical and modular algorithm combining optimization and complexity control approaches; MLRC,<sup>17</sup> a maximum likelihood method; PHD,<sup>18</sup> a neural network-based method; PREDATOR,<sup>19</sup> a method based on nearest neighbor detection; and SOPM,<sup>20</sup> a self-optimization secondary structure prediction method.

### Fold Recognition

Fold recognition also was performed by using three independent methods, ranked among the best predictors in CASP and CAFASP experiments.<sup>11,21</sup> The methods used are as follows: the GenThreader available at <http://insulin.brunel.ac.uk/psipred/>,<sup>22</sup> a fast and powerful fold recognition method that combines sequence information with pseudo-energies obtained from solvation and contact potentials, previously derived from known protein structures; HFR available through the BIOINGBU server at <http://www.cs.bgu.ac.il/~bioingbu/>,<sup>23</sup> a hybrid fold recognition method that collects results from five different threading programs, combining them along with evolutionary information from sequence analysis, in a search for the most consistent fold prediction among them; and the 3D-PSSM at <http://www.bmm.icnet.uk/~3dpssm/>,<sup>24</sup> a method that combines multiple-sequence profiles with

structure-based profiles which include solvation potentials derived from known structures and predicted secondary structures. The top 20 hits obtained from each fold recognition method were classified in the core fold as per the SCOP<sup>25</sup> and FSSP<sup>26</sup> databases. The frequency of occurrence of each fold was analyzed. The alignments were then classified into regions that matched or mismatched the consensus secondary structure, allowing specific SCOP classes to be selected as potential candidates for model building. Each template was analyzed for consistency with the biochemical data available for the H1 C-terminal domain.

### Model Building

A detailed structural analysis of the chosen template, HMG-box, was conducted. The C-terminal domain of histone H1d was aligned to the structure-based alignment of the different proteins containing the HMG-box, keeping the predicted secondary structural elements and the known functional residues in view. An initial model for the domain was constructed manually by using FRODO,<sup>27</sup> based on the backbone conformation of the HMG-box template and the alignment with H1d. The model was then improved as detailed in Results. The models were regularized by energy minimization. A steepest descent refinement of 100 steps using DISCOVER interfaced with Insight-II was used to achieve this. Visualization, manipulations, and analysis of various structures were performed by using Insight-II (Accelrys Inc.). Several experimentally derived structures of complexes of HMG-box with DNA are available in the Protein Data Bank and served as templates for modeling the protein-DNA complex. DNA was then treated as a rigid body and its position was optimized with respect to the protein. The data from extensive site-directed mutagenesis of H1d was mapped onto the model for testing the validity of the model.

## RESULTS

### Overview of the Sequence Analysis

Histone H1d belongs to a family of 186 members in the Pfam database<sup>28</sup> placed in the functional category of essential nucleosomal components with a clear structural and functional annotation for the globular domain but does not contain any information about the C-terminal domain. A search through the Interpro database<sup>29</sup> shows that rat H1d has 256 orthologs, but here again no suitable structural or functional template has been identified for the C-terminal domain. The SMART<sup>30</sup> and DART<sup>31</sup> tools recognize regions 34–109 and 110–219 as separate domains, which indeed correspond to the globular domain and the C-terminal domain (H1d\_C), respectively. The COGNITOR tool does not reveal any related COGs at 3 Be clades, within the COG database.<sup>32</sup>

Domain organization of rat H1d illustrated in Figure 1 shows the three octapeptide repeat units containing the S/TPKK motifs within H1d\_C, which were recently shown by site-directed mutagenesis studies to be important for DNA condensation.<sup>33,34</sup> Among these octapeptide repeats, the first two are in tandem and span amino acids 144–159

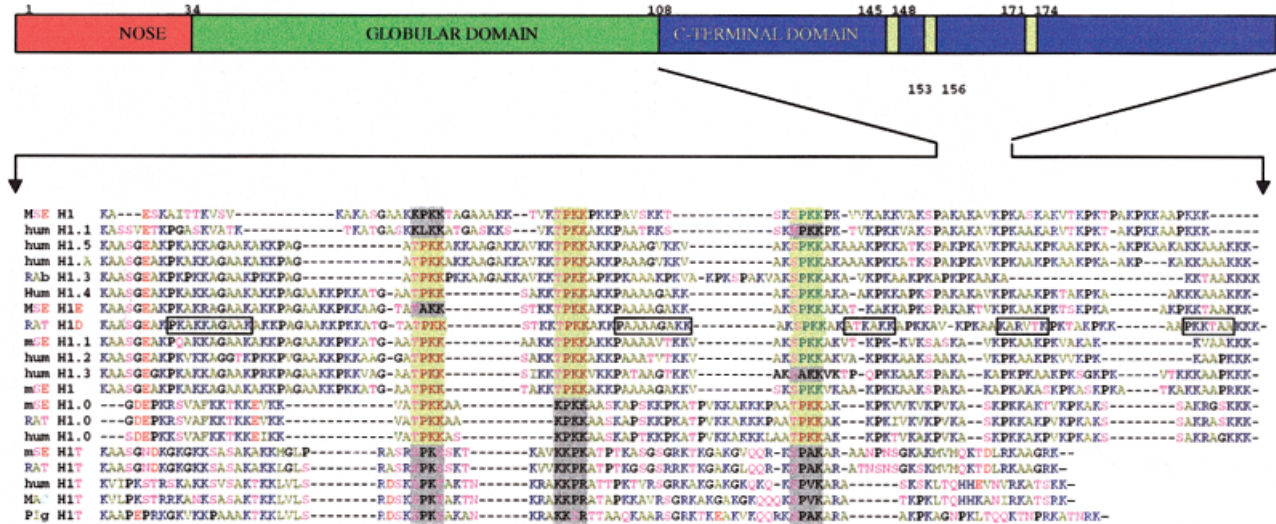


Fig. 1. Representation of the domain organization in histone H1d from rat (top). The three S/TPKK repeats within the C-terminal domain are highlighted by yellow bars. Residue numbers of these and of the domain boundaries are marked. Bottom: A multiple alignment of the C-terminal domains in several mammalian linker histones. Basic residues are shown in blue, acidic residues in red, polar uncharged residues in pink, prolines and glycines in black, and nonpolar residues are shown in gray. The three S/TPKK motifs are highlighted in yellow, and their variants observed in some sequences are highlighted in gray.



Fig. 2. A: Secondary structure prediction for the C-terminal domain of rat H1d, using different methods (see text). Helical regions are denoted by "h." The last line shows the consensus prediction obtained by using NPSA (15). B: Schematic representation of the three fingerprints observed for H1d\_C, positioned appropriately in context of its sequence. Prints corresponding to the high-mobility IY group proteins are shown as filled bars, those from the ATHOOK motif are shown as dot-filled bars, and those corresponding to the nonhistone HMG proteins are shown as striped bars.

with the proline residues being 146 and 154. The third octapeptide unit is present 10 amino acids away from the first two and spans amino acids 170–177 wherein the proline in the S/TPKK corresponds to amino acid number 172. A multiple alignment of the C-terminal domain of various histone H1 subtypes from several mammalian species also shown in Figure 1 highlights the conservation of the S/TPKK motifs despite diversity in the length of the domain.

**Structure Prediction**

The secondary structure predictions of H1d\_C using several methods consistently showed the presence of four

to five helices as illustrated in Figure 2(A). Differences in the secondary structure predictions from the different methods GORIV, HNNC, MLRC, PHD, PREDATOR, and SOPM were primarily limited to the boundaries of the helices. The consensus predictions were largely used to mark the boundaries. Where there was a conflict, a higher weight was given to the prediction reliability indices obtained through the "predictprotein" algorithm. Prediction of the structural class using both the sequence composition as well as through secondary structural elements unambiguously place the domain in an all- $\alpha$  structural class. Repeating the studies with other somatic H1 sequences gave consistent results. To minimize prediction

bias due to the higher percentage of lysines found in the sequence, the sequence was randomized, and the prediction was repeated with 20 such randomized sequences. The helical content and the position of the helices varied enormously in the randomized sequences, ranging from >99% helicity to <5% with the rest of the sequence in a random coil state, as against the 38% helical content in the original sequence. All of them, however, gave predominantly all- $\alpha$  predictions, suggesting that a composition as in the H1d\_C sequence would have a high propensity for forming all- $\alpha$  structures. The correct sequence, however, is important in predicting the positions and lengths of the individual helices.

The top hits from the various fold recognition methods conducted independently with the sequence of H1d\_C largely indicate an all- $\alpha$ -fold consistent with the predicted structural class and secondary structure. Analysis of their structures led us to classify them into six different classes as illustrated in Table I. These folds were consistently predicted for close homologues of H1d\_C. The regions within these structures that the H1d\_C was predicted to adopt invariably corresponded to three to four  $\alpha$ -helices, arranged approximately in an L-shaped structure in many of them. Many of these folds were also characteristic of protein families known to specifically bind to nucleic acids. This finding gains a particular relevance because H1d\_C also is known to interact with DNA, and the fold it adopts would obviously have to support such a function. These folds, although classified under separate SCOP classes, exhibit many common features in the regions relevant for this study (i.e., regions corresponding to those that align with the C-terminal domain sequence). The fact that three helices forming an approximately L-shaped structure has been predicted consistently suggests that to be the basic skeleton of the histone C-terminal domain structure. The folds of each of the structures among the top hits were analyzed by building models of H1d\_C, wherein its sequence was threaded onto each of the folds, based on the predicted alignments. These models were investigated in detail for (a) their compatibility with each template in the environment of the side-chains, (b) the overall alignments between the two sequences; (c) positions of the insertions and deletions with respect to the template structure, (d) the alignment of the observed secondary structural elements versus the predicted ones for the C-terminal domain, (e) solvent accessibility of the known functional residues, which in this case, are the three S/TPKK repeat units, and (f) the potential conservation of the key interactions between secondary structural elements in the template. This analysis resulted in identifying the HMG-box fold as the most appropriate candidate template for modeling the C-terminal domain.

Furthermore, evidence from a number of other analyses, discussed below, revealed a distant evolutionary link between H1d\_C and HMG-box proteins, thus strengthening the prediction of HMG-box fold in H1d\_C. (a) The finger print sequence patterns corresponding to the HMG and the related AT-hook proteins (PRINT HIGHMOBILITY, PR00930, *p* value of 1.9e-05; ATHOOK, PR00929, *p*

value of 2.4e-05; and NONHISHMG17, PR00925, *p* value of 3.1e-05), as defined in the PRINTS database, were detected in the sequence of H1d\_C [Fig. 2(B)]. (b) A distant sequence homology with the HMG-box protein (1aab) was also detected from a FASTA search (29.5% identity over a 78-amino acid stretch) of the Protein Data Bank. (c) The three functional motifs (octapeptide repeats) in the H1d\_C structural model occupy positions similar to the functional (DNA-binding) residues in the HMG-box proteins. (d) The HMG 1/2 protein family has two HMG domains present as tandem repeats within the protein,<sup>35</sup> resembling the architecture of somewhat similar domains within the linker histones, the globular domain, and the C-terminus. (The globular domain of histone H1 consisting of a three-helical bundle<sup>7</sup> was one of the lower scoring hits during structural homology searches.)

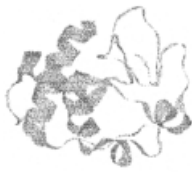


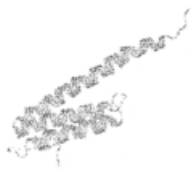


### HMG Proteins and Histone H1

An obvious question raised by the identification of an HMG-box fold in the C-terminus of H1d through fold recognition studies and the detection of an evolutionary link between the two is whether the functional profiles of the HMG proteins and H1 bear any similarities. Reviewing the information from the literature about various functional aspects of these proteins reveals significant similarities indeed, as outlined below, further strengthening the structure prediction. The high-mobility group (HMG) proteins are among the largest and best-characterized group of nonhistone chromosomal proteins. Although the structure of these chromosomal architecture proteins is well defined, their cellular function is not clearly understood. Linker histones, together with HMG proteins, are the major proteins that bind to the linker DNA in chromatin and exhibit both generalized and specific effects on gene transcription.<sup>36</sup> The binding of linker histones and HMG-1/2 to chromatin is believed to be competitive *in vivo*, although the mechanism is not clear. Similar to histone H1, HMG-1 has also been shown to protect DNA reconstituted into mononucleosomal<sup>37</sup> or dinucleosomal particles.<sup>38</sup> Both the proteins display similarities in many aspects of DNA binding. Both unwind DNA, and in most cases, both histone H1 and HMG-1 possess the same sequence and structure specificity. HMG-box proteins bind minor grooves of AT-rich DNA, as is the case with histone H1. Both bind to distorted helices such as those found in four-way junctions or cisplatin-modified sequences. HMG proteins significantly bend DNA as shown by circularization assays. Currently, it is believed that the linker histones H1 act as general repressors, whereas HMG1 and 2 act as transcriptional activators. The molecular basis of the modulation of transcription activity by these proteins is not clearly known but becomes comprehensible if they were to have similar structures. The competition between the HMG-proteins and histone H1 for binding to chromatin is also easily explained if they were to have similar structures.

### Model Building

An alignment derived from the structural superpositions of the various proteins containing the HMG-box fold

**TABLE I. List of the Folds Corresponding to the Top Hits From Different Prediction Methods**

PDB codes	Protein family/(ies)	Fold name (from SCOP)	Structure
1ccr, 1wad, 1jaf, 1cgo, 1a7v, 1cry	Cytochrome C	Cytochrome c, 3 helices, folded leaf, opened	
1aab, 1ckt, 1cg7, 1hsm, 1qrv, 2lef	HMG-1, HMG-D, Lymphoid enhancer protein	HMG-box, 3-helices, irregular array	
1flm	Outer surface glycoprotein C	Four-helical up-and-down bundle, closed or partly opened, left-handed twist; up-and-down	
1ff, 1hr0	Ribosomal protein, S20	Spectrin repeat-like 3 helices; bundle, closed, left-handed twist; up-and-down	
1bs2	t-RNA synthetase	Anticodon-binding domain of a subclass of class I aminoacyl-tRNA synthetases, 4 helices, bundle; one loop crosses over one side of the bundle	
1bxi	Colicin, DNase domain	His-Me finger endonucleases, segregated $\alpha$ - and $\beta$ -motifs, HNH motif	

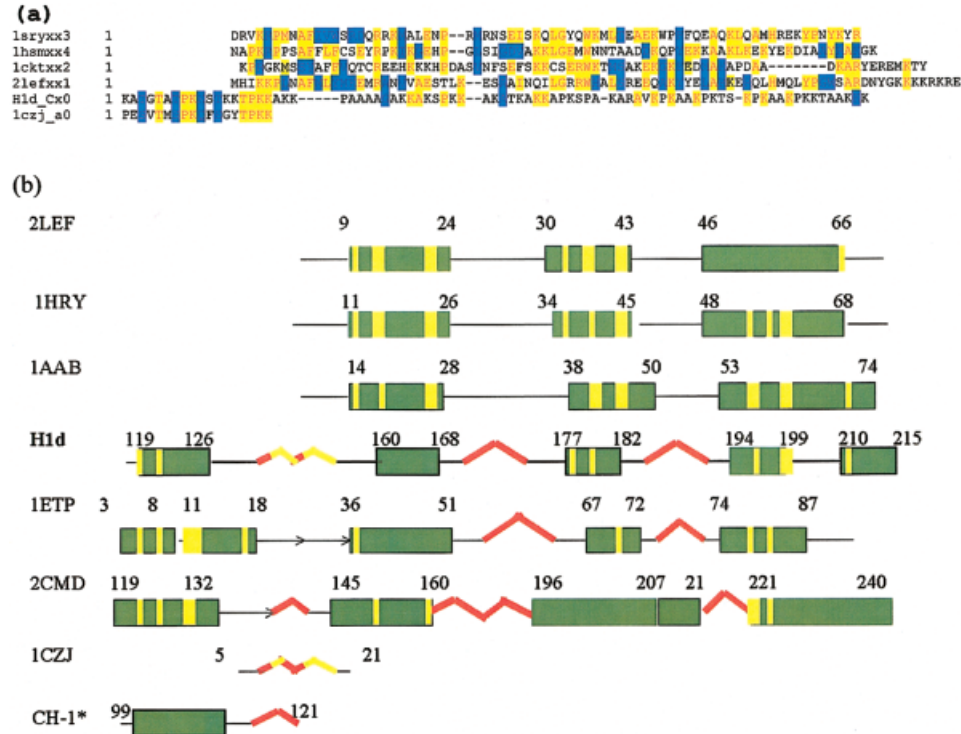


Fig. 3. **A:** Alignment of H1d\_C to the structure-based alignment of the HMG proteins 1SRY, 1HSM, 1CKT, and 2LEF. The figure was prepared by using BOXSHADE written by Hoffman and Baron, available at [http://www.ch.embnet.org/software/BOX\\_form.html](http://www.ch.embnet.org/software/BOX_form.html). The alignment of the turn regions of H1d\_C with the fragment from 1CZJ used for model building is also shown. Identical amino acids with a threshold of 0.4 (40% of the sequences) are shown in red and highlighted in yellow, whereas amino acids, which have conservative substitutions, are shown in blue and highlighted in cyan. **B:** Schematic representation of organization of secondary structural elements predicted for the H1d\_C. Helices are represented by green rectangles, and turns are indicated in red as inverted Vs. Above this, an alignment of this domain with a structural alignment of the different HMG-box proteins used for model building is shown schematically, indicated by their PDB codes on the left. The alignment with other structures that showed significant fold, sequence, and secondary structural compatibility is shown below along with their PDB codes. Residue numbers are also given for each structure to indicate the position of secondary structural elements. The figure is approximately to scale. Sequence similarities of the H1d\_C with the various structures shown are highlighted in yellow. CH-1 refers to the NMR structure of the N-terminal peptide of H1<sup>o</sup> (see text for details). The CH-1 peptide shows high-sequence similarity with rat H1d and has not been highlighted here for clarity.

revealed a high degree of structure conservation despite very low sequence conservation. An alignment of residues 151–218 of H1d\_C to the structure-based alignment of three of the HMG proteins, 1CKT, a crystal structure of HMG-1 protein complexed with cis-platin modified DNA and 2LEF, a solution structure of the lymphoid enhancer factor-1 in complex with DNA, 1SRY a solution structure of sex-determining hRY protein, whose sequences were sufficiently different for each is shown in Figure 3.

An initial model of H1d\_C was built on the basis of the above templates. The model includes the three helices forming the L-shape of the domain, the third S/TPKK motif, and provides clues for positioning the first two S/TPKK motifs, based on the position of the N-terminal basic segment in Lef-1, which interacts with DNA. NMR studies of the S/TPKK motifs have shown that a peptide corresponding to the sequence motif has a characteristic turn structure.<sup>39</sup>

A search through the PDB revealed a cytochrome C3 (1CZJ) to have a very similar sequence and structural motif. By taking advantage of the knowledge of the precise arrangement of atoms in this motif from several independent structural studies, this substructure was incorpo-

rated into the model at appropriate positions [Fig. 4(A)]. The position and conformation of the nucleic acid were derived from the structure of the Lef-1-DNA complex.<sup>40</sup> Although the HMG-box fold together with the turn regions account for the fold and function of most of the C-terminal domain, the first 31 residues still remained noninterpretable at this stage. Therefore, the possibility of extending the model at the N-terminus was explored by comparing with the other templates identified by fold recognition methods. Analyzing them in view of the sequence and overall structure compatibility (i.e., compatibility of the segment as part of an overall structure containing the HMG-box fold) resulted in recognizing cytochrome C4 (1ETP:residues 1–28) as a suitable candidate.

Further sequence and structural homology searches indicated a surprising structural similarity between a segment of malate dehydrogenase family [2CMD:147–232; see Fig. 3(B) with the HMG-box proteins (FSSP: Z-score 2.7, RMSD 3.8)]. Apart from that, residues 119–146 of this protein (2CMD) showed a considerable sequence and secondary structural compatibility with the N-terminal region of the H1d C-terminal domain. The structures of the segments from both these different proteins were reason-

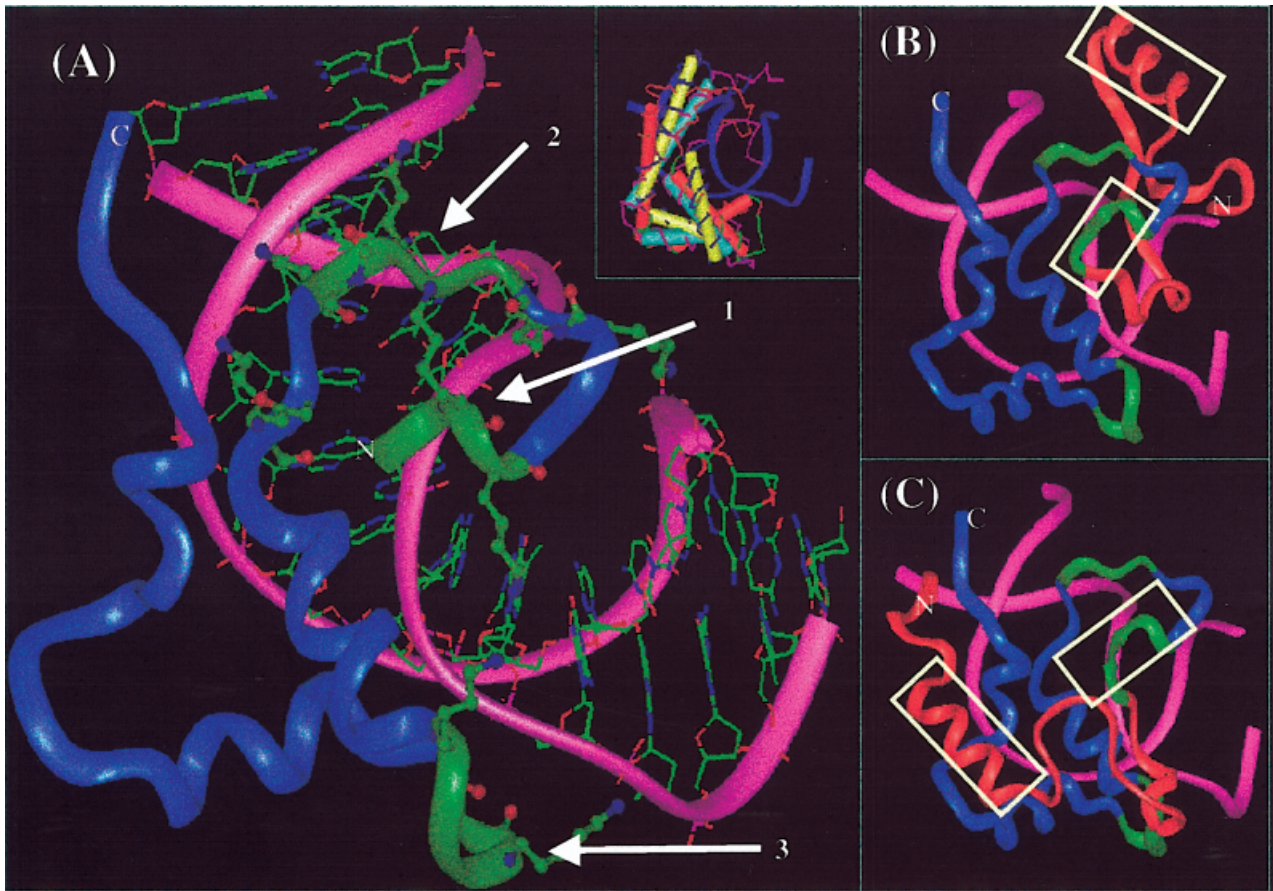


Fig. 4. **A:** Model of the C-terminal domain of rat histone H1d corresponding to the HMG-box domain and the DNA-condensing units. DNA docked into this model based on that of 2LEF is indicated by a violet ribbon. Superposition of the three HMG-box structures used for model building is shown in the inset (2LEF in red, 1HRY in yellow, and 1AAB in cyan). S/TPKK motifs are shown in green and labeled 1, 2, and 3, respectively, and the lysine side-chains in them are shown in ball-and-sticks representation. The N- and C-termini are marked. **B:** Full model of the C-terminal domain of H1 consisting of the HMG-box domain (yellow) along with its DNA condensing units (green) as well as the N-terminal segment of the C-terminal domain (red). The N-terminal segment modeled on the basis of malate dehydrogenase has been oriented to maximize interactions with DNA. The helix and the turn of the N-terminal segment corresponding to the CH-1 peptide are highlighted in pale yellow. **C:** Model similar to that in (B) except the orientation of the N-terminal segment (red), which has been modeled on the basis of the corresponding segment in cytochrome C4.

ably superimposable, consisting of either two  $\alpha$ -helices or one  $\alpha$ -helix followed by an extended region, from which a consensus template was generated, which matched with the secondary structure prediction of the H1d\_C.

However, positioning the segments thus built, with respect to the HMG-box domain, gave rise to two possibilities, one derived from the position of the segment in malate dehydrogenase and the other corresponding to that in cytochrome C4. Models corresponding to both orientations have been built as shown in Figures 4(B) and 4(C). In both cases, the model corresponding to the HMG-box domain of H1d\_C and the consensus template for the first part of the domain, oriented on the basis of either malate dehydrogenase or cytochrome C4, were stitched together to build the two alternate models. The interactions with DNA appear to be maximal in the first orientation, whereas the second orientation appears structurally more integral. It must be mentioned here that the HMG-box fold in HMG proteins is predominantly stabilized by an aromatic cluster of residues between the second and the third helices, a feature not observed in H1d\_C. However, it is clear from

the circular dichroism and other biophysical studies reported for H1d\_C<sup>5</sup> that its structure would get stabilized only on binding to DNA. Thus, interactions with DNA appear to be the predominant stabilization force for the H1d\_C structure.

In our second model in which the first 31 residues have been oriented on the basis of the cytochrome c structure, this segment interacts extensively with the second and the third helices of the L-shaped three-helical HMG-box, perhaps compensating for the absence of the aromatic cluster and rendering additional stability to the structure. Arguably, the structure of either cytochrome C4 or malate dehydrogenase, both distant structural neighbors of the HMG-box, could have been used as a template for the C-terminal domain. But it was reasoned that both these proteins, unlike the HMG proteins, are not DNA-binding proteins; the substructures corresponding to the HMG-box domain, although present may have altered to a considerable extent, especially in the three helices forming the DNA condensing unit.

Malate dehydrogenase has a considerable proportion of  $\beta$ -structures in addition to the  $\alpha$ -helices that showed

structural similarity, thus making it unsuitable to use as a template for the whole domain. Therefore, a hybrid model was constructed in which the amino acid residues from 109–139 of the C-terminus of histone H1 were built on the basis of either the CMD or ETP templates. The final models, which differed with each other only in the orientation of the N-terminal segments, were normalized for bond lengths and bond angles, and the nucleic acid docked into them by comparison with the LEF-1 structure.

### Experimental Support for the Model

(i) Correlation with circular dichroism data: CD studies have indicated the C-terminal domain to contain a significant proportion of  $\alpha$ -helical content only in the presence of trifluoroethanol, a commonly used inducer of helicity.<sup>5</sup> This seems to suggest that the domain adopts an ordered structure containing a few helices that requires either DNA or TFE for its formation or stability. Although this is no direct validation for the predicted structure, the presence of a HMG-box fold containing five helices, with strong interactions with DNA, is consistent with the spectroscopic observations and their implications.

(ii) Correlation with site-directed and deletion mutagenesis data: Earlier studies of mutating prolines to alanines in the three S/TPKK motifs from our laboratory<sup>33</sup> have shown the importance of first the motif in DNA binding and second the role of prolines in conferring the required substructures at the motifs. Subsequently, we conducted extensive deletion mutagenesis studies corresponding to the three motifs either singly or in combinations, the details of which will be reported elsewhere because it is outside the purview of this article.<sup>34</sup> The single mutants corresponding to the deletion of four amino acids of each motif show a reduction in DNA binding by 20%, the double mutants by 40%, and the triple mutant by 45%. Apart from these, a deletion mutant has been constructed in which a 34-residue segment corresponding to residues 144–177, housing all three S/TPKK motifs, have been deleted. This mutant shows 90% reduction in DNA binding and condensation, which led us to label this segment as the DNA-condensing unit. Our prediction depicts the 16mer and the 8mer sequences (corresponding to the three octapeptide repeats) as turns in the structure (Fig. 1). The striking feature that has to be considered here is that according to the mutational analysis, the 34 amino acid stretch within the C-terminus of histone H1, which binds and condenses DNA, corresponds exactly to this region. In addition, the concave surface of the L-shaped architecture of the HMG-box within the Lef-1 protein interacts with DNA.

(iii) Correlation with the experimental structure of the first 23 residues of the C-terminal domain: While these studies were in progress, Vila et al<sup>41</sup> studied the conformational properties of a peptide representing the sequence stretch from position 99–121 in the C-terminus of histone H1<sup>0</sup> (rat H1.0). This peptide sequence is immediately after the globular domain and just before the 34 amino acid stretch that we have identified in the present study as the DNA-condensing unit (see also Fig. 1). Their NMR analysis showed that although, in aqueous solution, the peptide

is unstructured, it obtained substantial  $\alpha$ -helicity in trifluoroethanol. The helical region revealed a strong amphipathic character, with all positively charged residues concentrated on one face of the helix and all the hydrophobic residues on the opposite face. They have extended these studies and have shown by Fourier transform infrared spectroscopy that DNA induces  $\alpha$ -helical segments similar to that observed in trifluoroethanol.<sup>42</sup> Thus, it is becoming increasingly obvious that the C-terminus of histone H1 has a sequence with a potential to adopt a fold containing regular secondary structures, which would also be necessary in determining the curvature and path of the linker DNA between two adjacent two nucleosomes. This peptide corresponds to residues 112–131 and residues 144–149 in the rat H1d sequence owing to the 12-residue insertion compared to H1<sup>0</sup>. It is gratifying to note that our model also shows a  $\alpha$ -helix and turn region for these two segments but is separated by an extended region because of the insertion. The helix corresponding to the CH-1 peptide in our model is highlighted in Figure 4(B) and (C).

(iv) Correlation with helicity measurements: The S/TPKK repeats in H1d\_C are separated by the spacer region, which has been consistently predicted as a helical segment. The helix after the 8mer ends in a SPAK sequence, which has been assigned to be a turn structure, and this distribution of the secondary structural elements within the overall fold consisting of alternating helices and turns aligns well with that of the HMG-box proteins. Along with the mutagenesis studies, we have shown that a deletion mutant in which the 10 amino acids between the second and the third motifs have been deleted shows a significant decrease in helicity content, whereas mutants of the S/TPKK motifs show an increase in the helicity values.<sup>34</sup> When analyzed in the light of modeling results, it becomes obvious why the removal of the spacer region ( $\Delta 10$  mutant) results in a decrease in  $\alpha$ -helicity (from 26.5% to 15.3%). The helicity changes are also seen in the C-terminal constructs. A further reduction in helicity was observed in the  $\Delta 34$  mutant (10.4%). On the contrary, deletion of the octapeptide repeats resulted in an increase in the induced helicity, suggesting that removal of the turn regions extends the helicity to the segment beyond the spacer region. These experimental data completely agree with the model described above for H1d\_C.

### Structure-Function Insights

It is clear from the results presented here that H1d\_C most probably adopts a conformation similar to the HMG-box domain present in rHMG 1, mLEF-1, and hSRY. It is worth discussing briefly the salient structural features of the interaction of LEF-1 and SRY with DNA. The structural data of Lef-1-DNA complex shows that Lef-1 wraps around and completely encompasses a highly distorted duplex DNA.<sup>40</sup> The central framework of the protein is formed by the characteristic L-shaped arrangement of the helices and an extended region seen previously in structures of HMG 1 and HMG D domains in the absence of DNA. The first two helices form one arm of the L, whereas the third helix and the extended N-terminal region, the



rest of the L. The DNA duplex binds to the concave surface of the Lef domain and is bent severely toward the major groove but retaining the Watson-Crick base pairing. The domain makes extensive and continuous contacts in the minor groove, which encompasses the entire region implicated in binding as evidenced by foot printing and mutagenesis studies. Bending and opening of the minor groove is accompanied by substantial narrowing and deepening of the major groove. Bending occurs throughout the nine base pair recognition sequence with an average total curvature summed up over the 15 base pair path of approximately  $117^\circ$  degrees. The SRY protein is a transcriptional activator of the Mullerian inhibiting substance (MIS) gene and is composed of three domains in which the central domain corresponds to a DNA-binding HMG-box. The solution structure of a specific complex of this HMG domain of SRY with a DNA octamer consisting of the MIS promoter has been solved.<sup>43</sup> It has a twisted letter "L" or a boomerang shape with irregular N-terminal and C-terminal strands that lie directly opposite to each other. The L-shape is generated through three helices in which the long arm of the "L" is formed by the third helix and the N-terminal strand, whereas the short arm of the "L" is formed by helices 1 and 2. On binding to hSRY-HMG domain, the DNA undergoes profound structural changes from B-type DNA in the free state to an underwound form that has features intermediate between A and B type DNA. The DNA in the complex is bent by  $\sim 70\text{--}80^\circ$ . The DNA is located in the concave surface of the L-shaped hSRY-HMG domain, and binding occurs exclusively in the minor groove of the DNA, causing widening of the groove. The conformation of the distorted DNA follows the contours of the concave binding surface perfectly, and the DNA is pushed away from the body of the protein. More recently, the crystal structure of HMG D-DNA complex was also solved.<sup>44</sup> Unlike LEF-1 and SRY, HMG D has minimal sequence specificity. It is surprising that the overall structure is very similar to that of LEF-1-DNA complex, and the DNA in the protein-DNA complex is bent by  $111^\circ$ .

An important correlation between the DNA condensation data and the model that is presented here is that the 34-amino acid stretch in the C-terminus of histone H1d, which we have identified as the DNA-condensing unit, corresponds structurally to the DNA-binding and -bending region of the HMG-box domains of Lef-1 and hSRY. The structural data show that the concave surface of the L-shaped molecular architecture of these proteins indeed interact with the DNA. Furthermore, the bent DNA is stabilized because of its position in the vicinity of the helix so that the lysine residues, which project from within the helix, interact with the DNA. The 34-amino acid stretch of the C-terminus of histone H1d corresponds to this region, and the helix aforementioned corresponds to the spacer region in between the 16mer and 8mer octapeptide containing the S/TPKK motifs.

The model predicted here for H1d\_C would require DNA to be bent for good binding, as seen in Lef-1 and SRY proteins. It appears that the 34-amino acid stretch encompassing the octapeptide repeats plays a key role in defining

the angle of bending with the octapeptide repeats containing the S/TPKK motifs functioning as the anchor points. It has been shown that the S/TPKK sequences can bind to DNA in the minor groove, and on bending there is a destabilization of the minor groove.<sup>45</sup> Because there are three octapeptide repeats in H1d\_C, it is reasonable to suggest that there would be destabilization of the minor groove on the binding of DNA by H1d\_C. This scenario is similar to that found in the HMG-box proteins. The bending might be the major contributing factor to the CD spectral observations of a progressive decrease in the positive ellipticity at 270 nm on interaction of DNA with histone H1.

Analysis of histone H1-DNA complexes using scanning force microscopy (SFM) has shown that in these globular complexes, the path of the DNA is not resolved. It seems from the general appearance that the DNA is bent and is probably around a histone H1 molecule.<sup>46</sup> Similar data have been obtained with the toroids formed by the interaction of isolated C-terminal domains of histone H1 with DNA.<sup>47</sup> Histone H1-induced DNA bending is physiologically relevant because the structure of the DNA obtained on interaction with histone H1 is close to the curvature of DNA in the condensed fiber. Some experiments have also indicated that compaction of linker DNA in nucleosomal templates is accompanied by bending or kinking of linker DNA.<sup>48</sup> The protein-induced DNA bending can be experimentally shown either by gel retardation assay with circularly permuted DNA substrate containing the target sequence for the sequence-specific-binding proteins.

The other method that has been used for the HMG proteins is the DNA circularization assay promoted by DNA ligase.<sup>49</sup> However, in the case of histone H1, both methods cannot be used because first it is not a sequence-specific binding protein and second it inhibits DNA ligase activity.<sup>50</sup> One of the predictions of correlation of the structural homology between the C-terminus of histone H1d and the HMG-box domains is that if the spacer of the 10 amino acids is removed, the entire structure collapses because of the absence of the helix and, therefore, the bent DNA duplex cannot be stabilized. In fact, this prediction is proved to be correct by the observation that the  $\Delta 10$  mutant loses 80% of the DNA condensation ability. It is important that this spacer has lysine residues.

It is worth noting that the octapeptide repeats are fairly conserved in most of the histone H1 subtypes across species except a few including the testis specific variant, H1t.<sup>51</sup> It is also interesting to note that their spacing and the length of the intervening sequence between the repeats vary among the different histone H1 subtypes (Fig. 1). This variation might contribute to the possible variation in the angle of bent DNA and also the extent of stabilization by the helical segment. It was shown in the DNA-binding studies that the  $\Delta 34$  mutant had residual DNA-binding activity, which might be due to regions beyond the 34 mer region. The structural data indicate that the last portion of the Lef-HMG-box beyond the third helix is a basic disordered segment, which interacts with DNA. In histone H1 also, there is a lysine-rich stretch

toward the end of the molecule, which might be contributing to residual DNA binding, similar to the scenario of LEF-1.

The S/TPKK motifs are known to recognize AT-rich sequences at the narrow minor grooves of the duplexes. We have shown earlier that a 16-mer peptide ATPKKSTKKT-PKKAKK, corresponding to the first and the second motifs found in H1d\_C, showed the highest preference for poly(dA-dT)-poly(dA-dT) containing SAR-DNA.<sup>52</sup> Our modeling studies clearly indicate that the first two motifs together bind at a minor groove, whereas the third motif appears to bind at the next groove. An investigation of genome sequences would be required to determine if there are indeed AT-rich segments that could indicate regions in linker DNA that could bind to the C-terminus of H1 and if they represent motifs correlating with the octapeptide repeats in H1.

The structural aspects related to the linker DNA has been a matter of controversy since the concept of higher order structure has emerged. It is still not clear what the path of linker DNA is. There have been suggestions that the path of the linker DNA influences the higher order structure and if solved will throw light on the folding pathway of polynucleosome fiber. In a recent study, Hamiche et al.<sup>53</sup> examined mononucleosomes reconstituted with the globular domain GH5 and also the full length of histone H5 containing the C-terminus by electron microscopy. Their electron micrographs clearly show that the globular domain of histone H5 increases the wrapping of DNA around the nucleosome core from 1.65–1.7 turns to 1.8–1.9 turns while at the same time the entering and exiting DNA are uncrossed. However, in the presence of full-length histone H5, a stalk is formed, which has been interpreted as the C-terminal tail bridging the entering and exiting DNA together to form a four stranded stem, which spans around a distance of 30 nm. The identification of the HMG domain in the C-terminus of histone H1 and the similarity of its structure with those of Lef-1 and SRY should stimulate efforts to understand the role of the C-terminus of histone H1 in chromatin folding. The structural homology of the C-terminus of histone H1d with HMG domains of transcription factors Lef-1 and SRY also represents yet another example of a structural connection that is emerging between the chromatin architectural proteins and transcription factors.<sup>54</sup>

## CONCLUSIONS

Sequence-based structure predictions are used routinely in structure-function studies of proteins, conducted almost automatically in many cases. However, predictions that use the existing knowledge on a particular protein family along with the recent advances of protein sequence and structural analyses have the potential to provide information stretching beyond the limits of entirely automated methods. This article shows an application of such an approach for deducing the structure of the C-terminal domain of histone H1. The prediction has facilitated the understanding of the mode of DNA binding by the C-terminal domain and provided an insight into the method

by which it achieves DNA condensation, which in turn influences chromatin folding. Identifying an HMG-box fold in the domain has also provided an explanation for many experimental observations such as competition in vivo between HMG proteins and H1 to bind to various target sites within chromatin, similarity in the DNA affinity profiles, and binding characteristics of HMG proteins with those of H1. It has also provided a structural basis to probe further the assembly of higher order structures of the chromatin fiber.

## ACKNOWLEDGMENTS

Use of facilities at the Interactive Graphics Based Molecular Modeling Facility and the Distributed Information Center (both supported by Department of Biotechnology), and the facilities at the Super Computer Education and Research Center are gratefully acknowledged.

## REFERENCES

1. Leuba SH, Yang G, Robert C, Somori B, van Holde K, Zlatanova J, Bustamante C. Three-dimensional structure of extended chromatin fibers as revealed by tapping mode scanning microscopy. *Proc Natl Acad Sci USA* 1994;91:11621–11625.
2. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosomal core particle at 2.8Å resolution. *Nature* 1997;389:351–360.
3. Clark DJ, Kimura T. Electrostatic mechanism of chromatin folding. *J Mol Biol* 1990;211:883–896.
4. Allan J, Cowling GJ, Harborne N, Cattani P, Craigie R, Gould H. Regulation of higher-order structure of chromatin by histones H1 and H5. *J Cell Biol* 1981;90:279–288.
5. Clark DJ, Hill CS, Martin SR, Thomas JO.  $\alpha$ -helix in the carboxy terminal domains of histone H1 and H5. *EMBO J* 1988;7:69–75.
6. Cerf C, Lippens G, Muyldermans S, Segers A, Ramakrishnan V, Wodak S, Halenga K, Wyns L. Homo- and heteronuclear two-dimensional NMR studies of the globular domain of histone H1: sequential assignment and secondary structure. *Biochemistry* 1993;32:11345–11351.
7. Ramakrishnan V, Finch JT, Graziano V, Lee PL, Sweet RM. Crystal structure of globular domain of histone H5 and its implications for nucleosome binding. *Nature* 1993;362:219–223.
8. Moran F, Montero F, Azorin J, Suau P. Condensation of DNA by the C-terminal domain of histone H1: a circular dichroism study. *Biophys Chem* 1985;22:125–129.
9. Shen X, Yu H, Weir JW, Gorovsky MA. Linker histones are not essential and affect chromatin compaction in vivo. *Cell* 1995;82:47–56.
10. Baker D, Sali A. Protein structure prediction and structural genomics. *Science* 2001;294:93–96.
11. Murzin AG. Progress in protein structure prediction. *Nat Struct Biol* 2001;8:110–112.
12. Pei J, Grishin NV. GGDEF domain is homologous to adenyl cyclase. *Proteins* 2001;42:210–216.
13. Thompson JD, Higgins DG, Gibson TJ. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighing, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994;22:4673–4680.
14. Attwood TK, Beck ME. PRINTS—a protein motif fingerprint database. *Protein Eng* 1994;7:841–848.
15. Combet C, Blanchet C, Geourjon C, Deléage G. NPS@: network protein sequence analysis. *Trends Biochem Sci* 2000;25:147–150.
16. Garnier J, Gibrat J-F, Robson B. GOR secondary structure prediction method version IV. *Methods Enzymol* 1996;266:540–553.
17. Guermeur Y, Geourjon C, Gallinari P, Deleage G. (MLRC): improved performance in protein secondary structure prediction by inhomogeneous score combination. *Bioinformatics* 1999;5:413–421.
18. Rost B. PHD: predicting one-dimensional protein structure by profile based neural networks. *Methods Enzymol* 1996;266:525–539.

19. Frishman D, Argos P. Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Eng* 1996;9:133–142.
20. Geourjon C, Deleage G. SOPM: a self-optimized method for protein secondary structure prediction. *Protein Eng* 1994;7:157–164.
21. Fischer D, Elofsson A, Rychlewski L. The 2000 Olympic Games of protein structure prediction; fully automated programs are being evaluated vis-a-vis human teams in the protein structure prediction experiment CAFASP2. *Protein Eng* 2000;13:667–670.
22. Jones DT, Miller RT, Thornton JM. Successful protein fold recognition by optimal sequence threading validated by rigorous blind testing. *Proteins* 1995;23:387–397.
23. Fischer D. Hybrid fold recognition: combining sequence derived properties with evolutionary information. *Pacific Symp Biocomput Hawaii* 2000;119–130.
24. Kelley LA, MacCallum RM, Sternberg MJE. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J Mol Biol* 2000;299:499–520.
25. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995;247:536–540.
26. Holm L, Sander C. Mapping the protein universe. *Science* 1996;273:595–603.
27. Jones TA. A graphics model building and refinement system for macromolecules. *J Appl Crystallogr* 1978;11:268–272.
28. Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer ELL. The Pfam protein families database *Nucleic Acids Res* 1999;27:260–262.
29. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MDR, Durbin R, Falquet L, Fleischmann W, Gouzy J, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A, Karavidopoulou Y, Lopez R, Marx B, Mulden NJ, Oinn TM, Pagni M, Servant F, Zdobnov EM. *Nucleic Acids Res* 2001;29:37–40.
30. Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting RC, Bork P. Recent improvements to the SMART domain-based sequence annotation resource *Nucleic Acids Res* 2002;30:242–244.
31. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res* 1997;25:3389–3402.
32. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* 2001;29:22–28.
33. Khadake JR, Rao MRS. Condensation of DNA and chromatin by an SPKK containing octapeptide repeat motif present in the C-terminus of Histone H1. *Biochemistry* 1997;36:1041–1051.
34. Bharath MMS, Sneha R, Chandra NR, Rao MRS. Identification of a 34 amino acid stretch within the C-terminus of histone H1 as the DNA condensing domain by site directed mutagenesis. *Biochemistry* 2002;41:7617–7627.
35. Bustin M, Reeves R. High mobility group chromosomal proteins: architectural components that facilitate chromatin function. *Prog Nucleic Acids Res Mol Biol* 1996;54:35–100.
36. Zlatanova J, van Holde KE. Linker histones versus HMG1/2: a struggle for dominance. *Bioessays* 1998;20:584–588.
37. Reeck GR, Isackson PJ, Teller DC. Domain structure in high molecular weight group non histone chromatin proteins. *Nature* 1982;300:76–78.
38. Onate SE, Prendergast P, Wagner JP, Nissen M, Reeves R, Pettijohn DE, Edwards, DP. The DNA-bending protein HMG-1 enhances progesterone receptor binding to its target DNA sequences. *Mol Cell Biol* 1994;14:3376–3391.
39. Suzuki M, Gerstein M, Johnson, T. An NMR study on the DNA binding SPKK motif and a model for its interaction with DNA. *Protein Eng* 1993;6:565–574.
40. Love JJ, Li X, Case DA, Giese K, Grosschedl R, Wright PE. Structural basis for DNA bending by the architectural transcription factor LEF1. *Nature* 1995;376:791–795.
41. Vila R, Ponte I, Jimenez MA, Rico M, Suau P. A helix-turn motif in the C- terminal domain of histone H1. *Protein Sci* 2000;9:627–636.
42. Vila R, Ponte I, Collado M, Arrondo JL, Suau P. Induction of secondary structure in A-COOH terminal peptide of histone H1 by interaction with DNA: an infrared spectroscopy. *J Biol Chem* 2001;276:30898–30903.
43. Werner MH, Herth R, Gronenhorst AM, Clore GM. Molecular basis of human 46 x,y sex reversal revealed from the three-dimensional solution structure of the human SRY-DNA complex. *Cell* 1995;81:705–714.
44. Murphy FV, Sweet RM, Churchill MEA. The structure of a chromosomal high mobility group protein DNA complex reveals sequence-neutral mechanisms important for non-sequence-specific DNA recognition. *EMBO J* 1999;18:6610–6618.
45. Takeuchi H, Sasamori J. Structural modification of DNA by a DNA-binding motif SPKK: detection of changes in base-pair hydrogen bonding and base stacking by UV-resonance Raman spectroscopy. *Biopolymers* 1995;35:359–367.
46. van Holde K, Zlatanova J. What determines the folding of chromatin fibre? *Proc Natl Acad Sci USA* 1996;93:10548–10555.
47. Butler PJ, Thomas, JD. Dinucleosomes show compaction by ionic strength, consistent with bending of linker DNA. *J Mol Biol* 1998;281:401–407.
48. Yao J, Lowary PT, Widom J. Direct detection of linker DNA bending in defined length oligomers of chromatin. *Proc Natl Acad Sci USA* 1990;87:7603–7607.
49. Payet D, Hillisch A, Lowe N, Diekmann S, Travers A. The recognition of distorted DNA structure by HMG-D: a foot printing and molecular modeling study. *J Mol Biol* 1999;294:79–91.
50. Ray E, Yaneva J, Ivanchenko M, van Holde K, Zlatanova J. Linker histones inhibit T4 and *Escherichia coli* DNA ligases. *Biochem Biophys Res Commun* 1996;222:512–518.
51. Khadake JR, Rao MRS. DNA and chromatin condensing properties of rat testes H1a and H1t compared to those of rat liver H1 bdec: H1t is a poor condenser of chromatin. *Biochemistry* 1995;34:15792–15801.
52. Khadake JR, Rao MRS. Preferential condensation of SAR-DNA by histone H1 and its SPKK containing octapeptide repeat motif. *FEBS Lett* 1997;400:183–186.
53. Hamiche A, Schultz P, Ramakrishnan V, Oudet P, Prunell A. Linker histone-dependent DNA structure in linear mononucleosomes. *J Mol Biol* 1996;257:30–42.
54. van Holde K, Zlatanova J. Chromatin architectural proteins and transcription factors: a structural connection. *Bioessays* 1996;18:697–700.