BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS **235,** 103–107 (1997)
ARTICLE NO. RC976748

Hydrophobic Cluster Analysis Reveals a Third Chromodomain in the *Tetrahymena* Pdd1p Protein of the Chromo Superfamily

Isabelle Callebaut,*[,1] Jean-Claude Courvalin,† Howard J. Worman,‡ and Jean-Paul Mornon*

**Systèmes Moléculaires et Biologie Structurale, LMCP, CNRS URA09, Universités Paris 6 et Paris 7, 4 Place Jussieu, 75252 Paris Cedex 05, France; †Département de Biologie Cellulaire, Institut Jacques Monod, CNRS, Université Paris 7, 2 Place Jussieu, 75251 Paris Cedex 05, France; and ‡Department of Medicine and Department of Anatomy and Cell Biology, College of Physicians and Surgeons, Columbia University, 630 West 168th Street, New York, New York 10032*

Received April 18, 1997

The protein Pdd1p (for programmed DNA degradation) links heterochromatin assembly and DNA elimination in *Tetrahymena* and has recently been identified as a new member of the chromo superfamily that contains two chromodomains. Using the sensitive bidimensional Hydrophobic Cluster Analysis (HCA) method, we have identified a third, highly divergent chromodomain in the Pdd1p sequence. Based on similarity to chromo shadow domains that mediate protein-protein interactions, this newly identified chromodomain may direct the binding of Pdd1p to other proteins. These findings suggest that Pdd1p may be the prototypical member of a new class in the chromo superfamily as all other members contain only one or two chromodomains. The results also demonstrate the power of HCA in identifying relationships which are not detected by conventional methods of sequence analysis.

The chromodomain (for *chr*omatin *o*rganization *mo*difier) is a conserved sequence of approximately 60 amino acids originally identified in the *Drosophila* proteins *h*eterochromatin *p*rotein 1 (HP1) and Polycomb (1). Many chromodomain-containing proteins have since been identified (2, 3), and based on the number and general features of the chromodomains, have been separated into three classes (2). Proteins with a classical chromodomain followed by a related but distinct sequence termed the chromo shadow domain have been termed class A, proteins with a single classical chromodomain have been termed class B and proteins with two classical chromodomains have been termed class C (2). Different classes of chromodomain proteins appear to have different functions, all of which are related to chromatin structure or organization. The class A protein HP1 is a suppressor of position effect variegation (4) and mammalian homologues of HP1 interact with transcriptional coactivators (5) and inner nuclear membrane protein LBR (6). The class B protein Polycomb is involved in the down-regulation of homeotic genes during development, presumably by stabilizing heterochromatin (1). Class C proteins, exemplified by mouse CHD-1, have additional DNA binding and helicase domains (7).

The protein Pdd1p that links heterochromatin assembly and DNA elimination in *Tetrahymena* has recently been described as a new member of the chromo superfamily with two chromodomains (8). However, only the first chromodomain of Pdd1p can be detected using the chromodomain profile described in PROSITE (PDOC00517) (9). This domain closely resembles canonical chromodomains although it does not fulfill the chromodomain consensus sequence. The second chromodomain in Pdd1p is one of the most divergent members of the family and, although clearly related to the first domain (28% sequence identity), it is not detected by the chromodomain signature or profile. In this respect, Pdd1p would not fall into the three above mentioned classes of chromodomain-containing proteins.

In the present study, we have used Hydrophobic Cluster Analysis (HCA) (10,11) to identify the presence of a divergent third chromo domain in Pdd1p, suggesting that it may be the prototypical member of a new class of proteins within the chromo superfamily. HCA is a sensitive method of sequence analysis that is able to detect 3-dimensional similarities between protein domains showing very limited relatedness. Its sen-

[1] Corresponding author. Systèmes Moléculaires et Biologie Structurale, Laboratoire de Minéralogie-Cristallographie, CNRS URA09, Universités Paris 6 et Paris 7, CASE 115, 4 Place Jussieu, 75252 Paris Cedex 05, France. Fax: 33-1-44 27 37 85. E-mail: callebau@lmcp.jussieu.fr.

0006-291X/97 $25.00

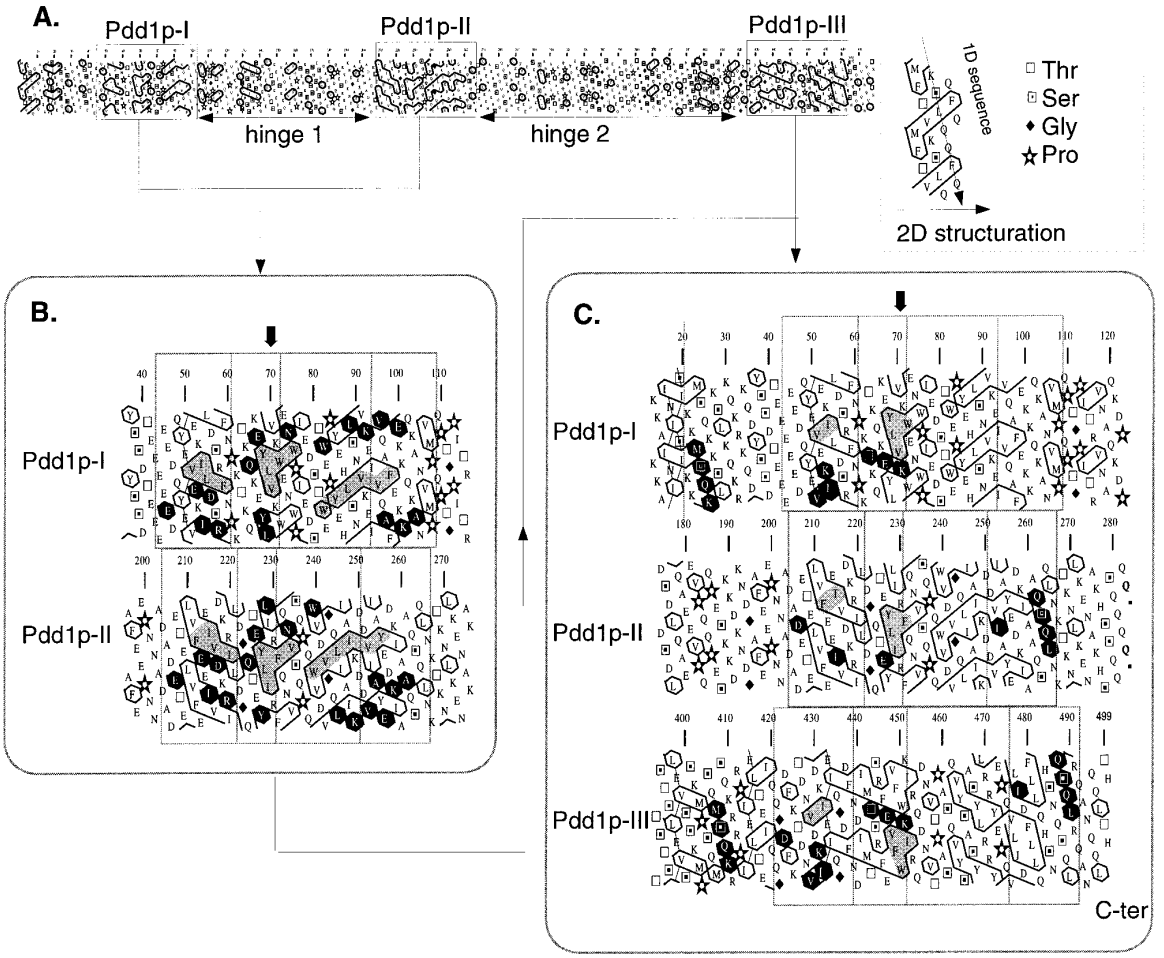

FIG. 1. Hydrophobic Cluster Analysis of the Pdd1p sequence (GenBank accession number U66364). (A) Global view of the Pdd1p HCA plot. Globular domains are boxed and the hinge regions separating them underlined. Symbols used in the plots and the manner in which the 1-dimensional sequence and 2-dimensional structuration are presented are indicated in the insert. (B) HCA comparison of the first and second globular domains of Pdd1p. Similar clusters are shown shaded; sequence identities are shown as white on a black background. The central cluster, which is highly conserved in the chromo superfamily (2), is indicated by an arrow. (C) HCA comparison of the three globular domains of Pdd1p.

sitivity at low levels of sequence identity, typically below the so-called "twilight zone" (25-30 %), stems from its ability to detect secondary structure elements (12). The effectiveness of the HCA method has been widely demonstrated (for examples see 13-18), especially in detecting internal repeated domains (15-18).

MATERIALS AND METHODS

Guidelines for the use of HCA have been published previously (10,11). Briefly, protein sequences are shown on a duplicated α-helical net with amino acid numbers indicated above. The contours of the hydrophobic residues are automatically drawn. The α-helical net has been shown to offer the best correspondence between the positions of hydrophobic clusters and regular secondary structures (12). Four symbols are used in the plots, regarding the specific structural behavior of the amino acids they represent: a star for proline, a diamond for glycine, a square for threonine and a dotted square for serine.

The statistical significances of alignments are assessed through the calculation of Z-scores. The Z-score represents difference between the alignment score under consideration and the mean score of a distribution computed for the alignment of sequence 1 versus a large number (10,000) of randomly shuffled versions of sequence 2 (13,16). These differences are expressed relative to the standard deviation of the random distribution, with values exceeding 3.0 considered to represent authentic relationships for small domains, as it is the case for the chromo superfamily (13,16).

RESULTS

Examination of the Pdd1p sequence using HCA demonstrates the presence of three globular domains separated by two hinge regions of about 90 (hinge 1) and 160 (hinge 2) amino acids, respectively (Figure 1A). Globular regions are characterized by a typical, thick distribution of hydrophobic clusters contrasting with hinge regions which only contain scattered hydropho-

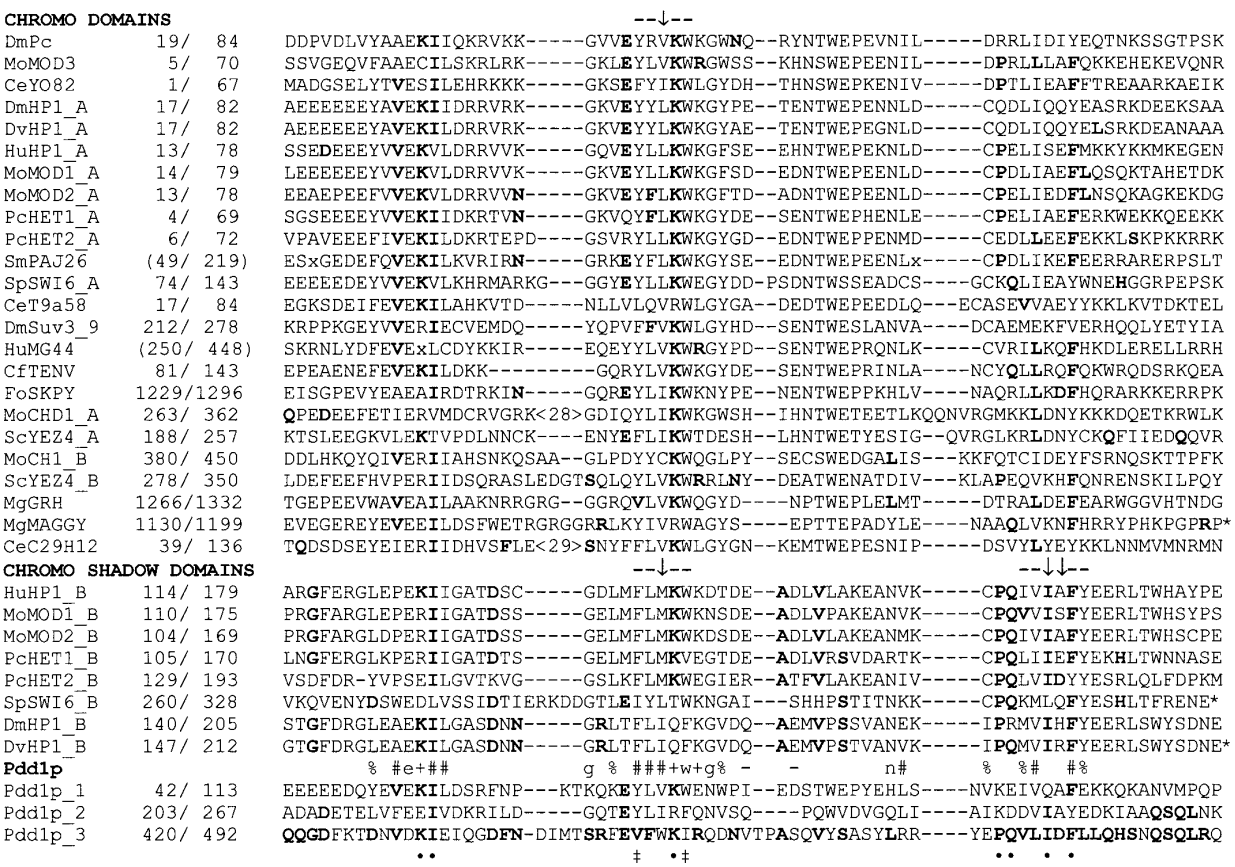

FIG. 2. One-dimensional sequence alignment of the Pdd1p chromodomains, deduced from HCA, with the other members of the chromo superfamily described in (2). The upper group of sequences contain classical chromodomains, the middle group chromo shadow domains and the lower group the Pdd1p chromo domains. Identical amino acids between Pdd1p-III and the other polypeptides are in bold print. ● indicates the positions of these identical amino acids which are highly conserved within the chromo superfamily. The domain positions in the protein sequences are indicated next to their names. x indicates ambiguous positions, * the end of protein sequences. Between the middle and lower groups, the consensus described by Aasland and Stewart (2) is shown (%, semi conserved hydrophobicity; #, strongly conserved hydrophobicity; −, conserved acidic residues; +, conserved basic residues). Note that he third chromo-like domain of Pdd1p meets the main features of this consensus. ‡ indicates the two positions in the conserved hydrophobic cluster where aromatic residues are substituted by aliphatic residues (V and I). The single and double arrows, as indicated in Figures 1 and 3, are also reported.

bic amino acids. The two chromo domains originally described by Madireddi *et al.* (8) correspond to the globular regions Pdd1p-I and Pdd1p-II (Figure 1B). These two domains share 28% sequence identity. The hydrophobic clusters YLVKW in Pdd1p-I and YLIRF in Pdd1p-II designated with an arrow in Figure 1B and strongly predicted to be a β-stand, correspond to the best conserved section between chromo domains (see Figure 2). This cluster can be coded 11101, where 1 is an hydrophobic amino acid and 0 any other one with the exception of proline which interrupts clusters.

The third globular domain of Pdd1p (Pdd1p-III), is found to have a similar length as the two preceding ones (approximately 60-70 amino acids). This globular domain contains several stretches of sequence conservation relative to the protein's first and second globular domains (Figure 1C). The hydrophobic pattern characterizing the "chromo fold" is also generally maintained in Pdd1p-III, with an absolute conservation of the typical cluster 11101 highlighted above (VFWKI, arrow in Figures 1B and 1C).

The first and fourth hydrophobic amino acids of the cluster characteristic of the "chromo fold", which in typical chromodomains are generally aromatic, are aliphatic in the third globular domain of Pdd1p (V and I, respectively; highlighted with the symbol ‡ in Figure 2), which may explain why this cluster escaped profile detection. Aliphatic amino acids are, however, also observed in these positions for some members of the chromo superfamily, for example a valine (MgGRH) or an isoleucine (SpSWI6 B) in the first position or a valine (PcHet1 B) in the fourth position (Figure 2). The three charged amino acids, E447, K451 and R453, which encircle the cluster 11101 in the third globular domain of Pdd1p, are also found to be well conserved in the chromo superfamily, especially K451 which rep-

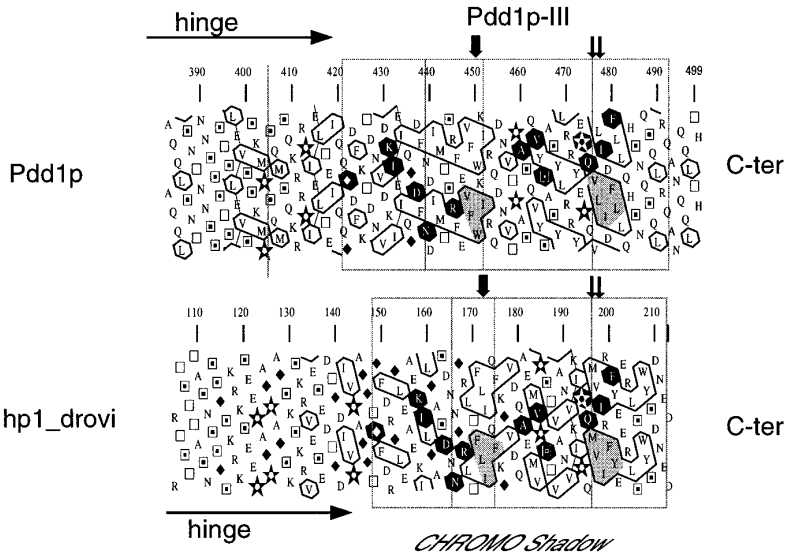

FIG. 3. HCA comparison of the third chromo domain of Pdd1p with the chromo shadow domain of *Drosophila virilis* (Swissprot P29227). Similar clusters are shown shaded, sequence identities are written white on a black background. The central cluster which is highly conserved in the chromo superfamily (2) is indicated by an arrow. The cluster representative of the chromo shadow family is indicated by a double arrow. Identity and similarity Z-scores calculated for this comparison are 4.8 and 4.2, respectively.

resents a prominent feature of the domain (Figure 2). Amino acids K431, I432, P474, Q475, I478 and F480 are also highly conserved in the chromo superfamily (highlighted with the symbol ● in Figure 2). Several cluster correspondences and sequence identities are particularly conserved between the third globular domain of Pdd1p and chromo shadow domains of other superfamily members (Figure 2).

A detailed comparison of the third globular domain of Pdd1p to the chromo shadow domain of *Drosophila virilis* HP1 is shown in Figure 3. The chromo shadow domain of *Drosophila virilis* HP1 shares 20% identity with the third chromodomain of Pdd1p (Figure 3). This percent identity is similar to that observed between the same chromo shadow domain and the classical chromodomain of the same protein (21%). In particular, the third part of the third globular domain of Pdd1p contains features typical of chromo shadow domains, as illustrated by the conservation of the cluster coded 111011 (double arrow in Figure 3) and by the sequence identities of P474, Q475, I478 and F480.

The statistical significance of the alignment of the third globular domain of Pdd1p with the eight chromo shadow domains shown in Figure 2 was assessed by the calculation of Z-scores. The mean of identity Z-scores is 3.7 (standard deviation 0.9) and the mean of similarity Z-scores (using BLOSUM62 matrix (19)) is 3.9 (standard deviation 0.8). These scores are consistent with a genuine relationship of this domain with the chromo shadow domains as values between 3 and 6 are frequently observed for divergent but related small domains (16). For comparison purposes, the identity and similarity Z-scores observed between the canonical chromodomains and the chromo shadow domains of the same eight proteins are 3.8 (σ 1.2) and 5.6 (σ 1.2) respectively, for a mean identity of 18.8 % (σ 3.9). These results are confirmed by profile-like procedures, based on the comparison of the third globular domain of Pdd1p not with isolated sequences but with the whole family (Callebaut *et al.*, unpublished results).

DISCUSSION

In this report, we predict that the Tetrahymena Pdd1p protein contains no other globular domains than three chromo-like domains that may mediate the functions of this protein. The length of the two hinge regions (~90 and 160 amino acids) is reminiscent of the structure of other members of the chromo superfamily, in particular those containing classical chromo and chromo shadow domains.

The third, heretofore undescribed chromodomain of Pdd1p (Pdd1p-III) resembles more chromo shadow domains than classical chromodomains and may function in mediating the interactions of Pdd1p with other chromodomains or chromatin-associated proteins, as reported for some chromo shadow domains. For example, the chromo shadow domains of HP1-type proteins is responsible for self-association (5, 20) as well as for the binding to either inner nuclear membrane protein LBR (20) or TIF transcriptional coactivators (5). A putative role of Pdd1-III in mediating protein-protein interactions therefore warrants investigation.

The third chromo-like domain of Pdd1p was detected using the HCA method but missed by methods such as BlastP (21) or consensus or profile detection specific to

the chromo superfamily (9). By putting sequence similarities in the context of 2-dimensional structuration, HCA adds a dimension to sequence analysis that overcomes the limitations of current 1-dimensional methods into and below the so-called "twilight zone" (typically under 25-30% sequence identity). The present data stress the power of the HCA method that may be particularly useful in interpreting genomic sequence data.

ACKNOWLEDGMENTS

This research was supported in part by a U.S.–France Cooperative Research grant from le Centre National de la Recherche Scientifique (J.-C.C.) and the National Science Foundation (H.J.W.). H.J.W. is an Irma T. Hirschl Scholar and this work was also supported in part by grants from the National Institutes of Health (CA66974) and the March of Dimes Birth Defects Foundation. We thank Dr. Paul Hossenlopp for helpful advice.

REFERENCES

1. Paro, R., and Hogness, D. S. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 263–267.
2. Aasland, R., and Stewart, A. F. (1995) *Nucleic Acids Res.* **23,** 3168–3173.
3. Koonin, E. V., Zhou, S., and Lucchesi, J. C. (1995) *Nucleic Acids Res.* **23,** 4229–4233.
4. Eissenberg, J. C., James, T. C., Foster-Hartnett, D. M., Hartnett, T., Ngan, V., and Elgin, S. C. R. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 9923–9927.
5. Le Douarin, B., Nielsen, A. L., Garnier, J.-M., Ichinose, H., Jeanmougin, F., Losson, R., and Chambon, P. (1996) *EMBO J.* **15,** 6701–6715.
6. Ye, Q., and Worman, H. J. (1996) *J. Biol. Chem.* **271,** 146543–14656.
7. Delmas, V., Stokes, D. G., and Perry, R. P. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 2414–2418.
8. Madireddi, M. T., Coyne, R. S., Smothers, J. F., Mickey, K. M., Yao, M.-C., and Allis, C. D. (1996) *Cell* **87,** 75–84.
9. Bairoch, A., Bucher, P., and Hoffmann, K. (1996) *Nucleic Acids Res.* **24,** 189–196.
10. Gaboriaud, C., Bissery, V., Benchetrit, T., and Mornon, J.-P. (1987) *FEBS Lett.* **224,** 149–155.
11. Lemesle-Varloot, L., Henrissat, B., Gaboriaud, C., Bissery, V., Morgat, A., and Mornon, J.-P. (1990) *Biochimie* **72,** 555–574.
12. Woodcock, S., Mornon, J.-P., and Henrissat, B. (1992) *Protein Eng.* **5,** 629–635.
13. Callebaut, I., and Mornon, J.-P. (1995) *FEBS Lett.* **374,** 211–215.
14. Henrissat, B., Callebaut, I., Fabrega, S., Lehn, P., Mornon, J.-P., and Davies, G. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 7090–7094.
15. Thoreau, E., Petridou, B., Kelly, P.A., Djiane, J., and Mornon, J.-P. (1991) *FEBS Lett.* **282,** 26–31.
16. Callebaut, I., and Mornon, J.-P. (1997) *FEBS Lett.* **400,** 25–30.
17. Callebaut, I., and Mornon, J.-P. (1997) *Biochem. J.* **321,** 125–132.
18. Callebaut, I., Renoir, J. M., Lebeau, M.-C., Massol, N., Burny, A., Baulieu, E. E., and Mornon, J.-P. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 6270–6274.
19. Henikoff, S., and Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 10915–10919.
20. Ye, Q., Callebaut, I., Courvalin, J.-C., and Worman, H. J. (1997) *J. Biol. Chem.*, in press.
21. Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990) *J. Mol. Biol.* **215,** 403–410.