Published in final edited form as:

Proteins. 2016 March ; 84(3): 397–407. doi:10.1002/prot.24986.

Identification of family-determining residues in Jumonji-C lysine demethylases: a sequence-based, family-wide classification

Patrick Slama

Independent researcher, Paris, France; formerly at Institute for Computational Medicine, the Johns Hopkins University, Baltimore MD

Abstract

Histone post-translational modifications play a critical role in the regulation of gene expression. Methylation of lysines at N-terminal tails of histones has been shown to be involved in such regulation. While this modification was long considered to be irreversible, two different classes of enzymes capable of carrying out the demethylation of histone lysines were recently identified: the oxidases, such as LSD1, and the oxygenases (JmjC-containing). Here, a family-wide analysis of the second of these classes is proposed, with over 300 proteins studied at the sequence level. We show that a correlated evolution analysis yields some position/residue pairs which are critical at comparing JmjC sequences and enables the classification of JmjC domains into five families. A few positions appear more frequently among conditions, such as positions 23 (directly C-terminal to the 2nd iron ligand), 24, 252 and 253. Implications of family conditions are studied in detail on PHF2, revealing the meaningfulness of the sequence-derived conditions at the structural level. These results should help obtain insights on the diversity of JmjC-containing proteins solely by considering some of the amino-acids present in their JmjC domain.

Keywords

Jumonji C; histone; lysine demethylase; subfamily; relative entropy; protein classification

INTRODUCTION

The DNA of eukaryotic cells is not readily accessible for transcription, but packed in an elaborate structure. The double-stranded DNA molecule first wraps around an octamer of histone proteins to form nucleosomes, which then pack more or less densely on top of each other to form chromatin. The packing of chromatin can be altered by the binding of chromatin remodeling complexes to the nucleosomes, and its unpacking results in a structure where access to DNA and histone tails is favored. Over the last decade, it has been shown that the N-terminal tails of histones, which protrude out of the nucleosome, are subject to multiple post-translational modifications [1; 2]. The consequences of these modifications are an alteration of the stability of the nucleosome, and a favored access to proteins or proteins complexes that specifically recognize these modifications.

Author contact: slama.p@gmail.com. Phone: +33-6-2832-1935.

Conflicts of Interest. None declared.

Apart from ubiquitylation [3], phosphorylation [4] or sumoylation [5], two further modifications, which involve smaller chemical groups, occur on histone tails: arginine and lysine methylation and lysine acetylation [6]. Lysine acetylation, as well as its reverse process, deacetylation, are currently thoroughly studied due to their likely involvement in multiple diseases, including some cancers [7; 8]. Methylation of both lysine and arginine can occur through similar mechanisms, which involve the transfer of a methyl group from S-adenosyl methionine to a side-chain nitrogen [9]. Various lysine residues from histones can be subject to either mono-, di- or tri-methylation, with evidences currently existing for Lys4 [10], Lys9, Lys27 [11; 12], Lys36 and Lys 79 from Histone 3 (HIS3) and Lys20 from Histone 4 (HIS4) [13; 14].

Removal of the N-methyl tag on histones is carried out by two different classes of enzymes. The LSD1/KDM1 demethylases are oxidases, which accept both mono- and di-methylated lysines as substrates [15]. The JumonjiC-domain containing demethylases are a broad family of enzymes, which were first evidenced on HeLa cells [16]. They have since then been shown to exist in various organisms, from e.g. yeast [17] to human [18; 19]. These two classes of enzymes proceed through different mechanisms: the first ones use the combination of an oxidase domain, a SWIRM domain and a tower domain to remove methyl groups from lysines in a FAD-dependent mechanism [19], while the second ones have an iron-containing active site that break up a dioxygen molecule and process methylated lysines in an α -ketoglutarate-dependent mechanism [6].

Many JmjC domain-containing proteins have been characterized in recent years, which display different specificities regarding both the histone lysine(s) they modify and the number of methyl groups they can tolerate on their substrate [16; 20; 21]. Due to the diversity of the domain composition of JmjC-containing demethylases, a classification of these domains would provide a better understanding of the causes for such diversity. Some classifications, which were based on sequence homology and similarities in surrounding domains, have already been proposed [22; 23; 24]. Still, a detailed analysis of these domains at the sequence level is missing.

Here, using proteins that were recently described and their respective substrate specificities, we propose a more complete analysis of the JmjC-containing demethylases sequences, in a study that involves more than 300 non-redundant protein sequences. A residue-residue co-evolution analysis [25; 26] is performed on JmjC domains (see Figure 1), which relies on a method recently elaborated on PHD fingers [27]. Some important position-residue pairs (alignment “conditions”) could be highlighted, with multiple conditions arising at positions 23 (directly C-terminal to the second iron ligand), 252 and 253 (respectively, Asn-2 and Asn-1 in our alignment). The grouping of JmjC domains according to these highlighted residues is described. Calculations on three-dimensional structures highlight the role of the positions evidenced by the sequence analysis within JmjC domains. Altogether, the results presented here should facilitate, through the simple reading of some specific positions in their sequences, the functional assignment of new JmjC proteins.

METHODS

JmjC domains sequence alignment

An alignment of JmjC domains was downloaded from the pfam database, version 24.0 [28]. The alignment was filtered for sequences possessing iron ligands (His/Tyr and Asp/Glu respectively) and manually refined, which produced 847 sequences. This new alignment was manually edited for obsolete sequences at time of calculation and submitted to a sequence identity-based clustering by software CD-hit [29], which yielded 363 distinct sequences. Residues were numbered from the most N-terminal residue of the alignment over all proteins to the most C-terminal residue, which had a position number of 256. The iron-binding residues respectively corresponded to positions 20 (His¹), 22 (Asp/Glu, later shortcut as 'Glu') and 232 (His²). Some positions with strong conservation trends in our alignment, such as position 6 which possesses a hydrophobic residue in 92% of the sequences, and position 254, which has an Asn in 89% of the sequences, will be referred to as 'Phi' and 'Asn', respectively. Alignment positions given relative to these positions in the text may not correspond strictly to the position numbering for a given protein and will rely only on the number of residues between two positions in the majority of the sequences, i.e. gaps present in 1 or 2% of the sequences will not be considered.

Selection of sequence 'conditions' and clustering of the conditions

The complete process of the selection of the most significant conditions, their comparison to each other and their clustering into different families was described earlier on PHD fingers and was further validated on chromo-domains and on acyltransferases [27]. Briefly, 110 observation positions were selected in the JmjC alignment, as the positions which had less than 30% gaps in our 363 sequences alignment. All possible amino-acid types which were observed at least 25 times at each position were used as conditions. The coupling of these conditions (e.g. $j:b$ for 'residue type b at position j ') to all other observation positions i (Figure 1) was measured by comparing the overall distribution in the JmjC alignment $p_i(a)$ to the conditional distribution $p_i(a|j:b)$ using relative entropy:

$$KL_i(j:b) = \sum_a p_i(a|j:b) \times \log \frac{p_i(a|j:b)}{p_i(a)} \quad (\text{Equation 1})$$

All resulting coupling values were ordered by decreasing values, and all conditions with at least one coupling in the first percentile of coupling values were preserved. The resulting set of conditions was submitted to affinity clustering, by using a symmetrized version of the relative entropy as a dissimilarity measure to compare any pair of conditions $j:b$ and $k:c$ (Equation 2; 'obs' represents the set of observation positions). Affinity clustering was preferred to other clustering methods as producing more reproducible cluster centroids (see [24]). This produced clusters or families of conditions which had similar KL values to the observation positions. The clustering process was also applied to conditions defined on other percentile filtering on coupling values (0.25 % to 2 %).

$$D_{KL}(j:b, k:c) = \frac{1}{|obs|} \sum_{\substack{i \in obs \\ i \notin j, k}} \frac{1}{2} [KL_i(j:b | k:c) + KL_i(k:c | j:b)] \quad (\text{Equation 2})$$

Assignment of a sequence to a family of conditions

Each sequence from the alignment was analyzed for the highlighted conditions from each family it fulfilled. A sequence was assigned to the family for which the number of conditions fulfilled was highest. In cases where two families corresponded to the highest number of conditions, no assignment was produced. ‘Uniquely assigned’ proteins or strict members were those that fulfilled conditions from only one family, i.e. which had only one non-null count of conditions.

Structure prediction

Structures were determined for JmjC domains using existing structures and homology modeling software Modeller (release 9v7 [30]). A mutated structure for PHF2 was generated using the structure for PHF2 (PDB entry 3PU8) and that for PHF8 (entry 3KV4) as templates, with two positions from PHF2 mutated according to residues from PHF8 (see text).

A structure for human JmjD4 was predicted using the structures for PHF8, for JmjD2A (PDB entry 2P5B) and for JmjD6 (PDB entry 3K2O) as templates, which presented respective sequence identities of 20%, 22% and 38% to the catalytic core of JmjD4, respectively. Three models were generated, two of which were considered for analysis: that with highest structural similarity to PHF8 (RMSD 2.18 Å) and that with lowest similarity to PHF8 (RMSD 3.48 Å). A Fe ion was incorporated in the models at the barycenter of N from His235 and His315 and C from Asp237. Docking was performed using FlexX (BioSolveIT GmbH, Sankt Augustin, Germany), without using the ‘access scaling’ option. Only the first pose of the docking calculation was considered, i.e. the one with the best overall score.

Active site detection

An active-site likelihood score was calculated using histone-free structures for KDM7A (3N9M), PHF8 (2WWU), PHF2 (3PU8), JmjD2A (2GP5) and JmjD6 (3K2O), as well as the predicted structure for JmjD4 (this study). The score was a combination of residue-residue network parameters and residue type, as previously described [31]. When using the most conservative threshold (*MDevI*), only one residue had a score above that threshold for the JmjD4 structure. For the sake of comparison, and in order to include residues other than the iron ligands in the comparison, we chose to consider the five residues with highest score for the JmjD4 model structure. This number corresponded to the average of the number of detected active site residues over the other proteins analyzed.

RESULTS

Families of conditions

A sequence alignment of 363 non-redundant JmjC domains (see METHODS for the selection of the sequences) was submitted to a residue-residue correlation analysis. Relative entropy was used as a correlation measure (Equation 1), as recently described [27]. Briefly, a correlation between two alignment positions i and j was measured by comparing the overall amino-acid distribution at position i to that obtained at position i when considering the subset of sequences where residue at position j was fixed to a given type; fixing a residue type at an alignment position was coined a 'condition'. Couplings between alignment positions and conditions (a condition is an alignment position:residue type pair) were ranked in decreasing relative entropy values (Figure 1). High relative entropy values for $KLi(j:b)$ imply that condition $j:b$ (see Figure 1.) significantly alters the residue distribution at alignment position i . The set of conditions having at least one coupling value in the first percentile of coupling values, the 'first-percentile conditions', included fifty-three conditions (Supplementary table S1). When using affinity clustering to compare the resulting disparity values, these conditions clustered into five families (see METHODS).

Some positions will be referred to relative to the most conserved alignment positions, such as 'Glu' for position 22 (the second iron ligand, which is conserved as a Glu or Asp in our whole alignment) or 'Asn' for position 254, which is conserved in most sequences (see METHODS for further details).

The five families of conditions are described in Table 1, with centroid conditions in the top row for each family. Families I and IV are respectively composed of 14 and 15 conditions, while family V is made up of as few as 4 conditions. Conditions originating at positions 4 (Phi-2), 23 (Glu+1), 252 (Asn-2) and 253 (Asn-1) appear in multiple clusters. For the latter, discrepancies in the role of these positions are expected between family I and family IV with respect to residue packing and hydrophobic contacts, as the first family includes condition 253:G, while position 253 is present in the second family under condition 253:W. Both clusters III and IV include two distinct conditions originating at the same positions, respectively 24 (His¹+4) and 230 (His²-2). It can also be noted that conditions at positions 252 and 253 often appear simultaneously in a given family. These trends are discussed below.

Assignment of JmjC sequences to families of conditions

The sequence alignment of JmjC domains was analyzed with respect to the clusters of conditions obtained. By considering the different 'conditions' from each family and whether a sequence satisfies mainly conditions from one family or another, one can translate the families of conditions into families of proteins. Assignment of a JmjC sequence to a family of conditions (see METHODS for details) was possible for 304 domains (84 % of the domains analyzed), with 234 sequences satisfying conditions from a single family (65 %, Table 2). Family assignment showed that the number of proteins that belonged to a family was not correlated with the number of conditions that defined that family, as e.g. seen when comparing the sizes of family I and V (Table 2). Moreover, overlap between families was

not related to the number of proteins that were assigned to each family. Family III, though the second biggest in number of proteins, had e.g. 93 % of its members strictly assigned to it.

Family I includes KDM7, which satisfies eleven of the family's fourteen conditions, and none from any other family. Proteins JHD1, PHF2 and PHF8, which have a PHD finger N-terminal to their JmjC domain, like KDM7, are also strict members of this family (Supplementary Table 2 and Table 2). JmjD4 proteins belong to family II, with e.g. all conditions from this family fulfilled by the chicken protein. Protein KDM4D and the KDM4D-like demethylase are members of family III (Figure 2), as well as e.g. protein RPH1 from yeast. KDM6B (JmjD3) fulfilled 12 of the conditions that define family IV. Finally, JmjD6 and PTDSR belong to family V, with again no conditions from other families fulfilled. Overall, the method produced a well-discriminated classification of JmjC domains, with little overlap between families, and multiple proteins as strict members of a family. Family III has an almost unique set of condition-defining positions, with e.g. no condition at positions 23, 252 or 253, but conditions at positions that only appear in this family (Table 1).

The validity of the family classification was asserted by comparing it to a more traditional clustering method, as obtained using an amino-acid similarity matrix over the whole alignment. For all of the sequences shown in Figure 2, our condition-based clustering was in agreement with the identity-based clustering (Figure 3). This suggested that the first-percentile conditions defined on JmjC domains were sufficient for comparing their sequences, i.e. that the sequence information contained in these conditions were efficiently summarizing the whole sequence of the core of the JmjC domain. Regarding more distant similarities, families II (JmjD4) and V (JmjD6) were classified as the most similar based on the amino-acid similarity matrix, while the conditions classification suggests that three positions (23, 252 and 253) could be sufficient to differentiate most of their respective members (Table 1).

Among the protein sequences that fulfilled no condition, one was from *Arabidopsis thaliana* (Q94BQ7, Putative RB-binding protein, Figure 2), one from *Xenopus laevis* (A4QNS2) and one from *Xenopus tropicalis* (A0JM02, aka KDM5C). Due to their identical residues at multiple positions, among which some where first-percentile conditions originate, the first and last of these proteins are likely to be classified in a same family when using a classification including more conditions. Indeed, when using the second-percentile conditions for defining the families (88 conditions as opposed to 53 first-percentile conditions), these two proteins belonged to the same family of conditions (Supplementary table S3). Still, the first-percentile classification was preferred overall, as producing a better combined score on sensitivity and strict membership (Table 2). A whole-sequence based clustering also placed Q94BQ7 in the same main branch as KDM4D proteins and KDM5 (Figure 3).

A structural role for the conditions was next searched among JmjC domains, first in family I, in order to address controversial results on PHF2, next by studying possible interactions of JmjD4, a member of family II, with histones. peptides.

Structural implications of conditions from family I: PHF2 vs PHF8

PHF2 has been shown to bind HIS3-Lys4Me₃ through its PHD finger, while an *in vivo* activity for its demethylase domain could be observed on HIS3-Lys9Me [32]. Still, later *in vitro* experiments contradict these results [33]. In order to address the homogeneity of family I and to bring more elements to the debate of a possible lysine-demethylase activity for PHF2, we analyzed its sequence discrepancies to other proteins from the family, and the influence these could have on histone recognition. Specificities in the sequence of PHF2 are the presence of a tyrosine as the third iron ligand (His²), that of a Ser at position 23 (Glu+1) and that of a His at position 254. At position 23, a Phe is present, which interacts with the methyl group from HIS3-Thr11 (Table 3). At position 254, a conserved Asn is observed in most JmjC domains (Figure 2). PHF8 and KDM7, which were assigned to the same JmjC domain family, both have a Phe at position 23, with 23:F a condition from family I (Table 1). The presence of a His at position 254, in a cavity where the methylated lysine of the substrate is normally present, as observed for PHF8 with HIS3-LysMe₂ [34] and for JmjD2A [35], might hinder the access of the substrate ammonium group to this cavity, due to charge repulsions. As mutation of Tyr at His² was shown not to affect the activity of PHF2 [33], we analyzed the structural implications of the sequence differences of PHF2 at positions 23 and 254. A structure for PHF2 with mutations Ser Phe* at position 23 (Glu+1) and His Asn* at position 254 was thus predicted (* designates a mutated residue in our hypothetical PHF2 sequence).

Superimposition of this hypothetical PHF2* structure to those of PHF2 and PHF8 suggested that Phe* at position 23 would provide a docking site for the alkyl chain of a lysine residue, while the side-chain of Asn* at position 254 would have the same orientation as that of His in wild-type PHF2 (Supplementary figure S4). Asn* would also develop a hydrogen bond with the iron-binding Asp at position 22. This analysis thus underlines the importance of condition 23:F from family I and that of having a His at position 254, which may justify the absence of a demethylase activity for PHF2 *in vitro* [33].

Structural implications of family definitions: comparison of active sites

Further common and specific trends across the families of JmjC proteins were obtained by analyzing their active sites. Additional structural information for the families of JmjC domain-containing proteins for which no histone-docked structure exists (family II and family V) was obtained using an active-site detection algorithm that relies on three-dimensional structures [31]. This algorithm was chosen due to the high accuracy it produces on functional sites. The detection was performed on JmjD4 and JmjD6, which respectively belong to these two families, along with structures from proteins belonging to family I and family III, for which docking is described in the literature [33; 36]. For JmjD4, a model structure was used (see METHODS).

For all of these proteins, two to three of the iron-binding residues were among the residues with highest scores (Table 4), which proved the relevance of the score on JmjC domains. Moreover, in four out of the five proteins analyzed, Lys at position 77 was detected as an active site residue. In PHF8, this residue (Lys264) had the fifth highest score, and was thus the first one below the *MDev1* threshold. In JmjD6, Mantri and Schofield recently showed

that mutation of the residue at the corresponding position (Lys204) abolished hydroxylation of peptide Luc7like2 267–278, and proposed that it played a role in 2-oxoglutarate binding [37]. In family I, for PHF2, as for PHF8 and KDM7, the residue with highest score was the second iron ligand, an Asp residue (Table 4). The prediction for PHF2 showed differences with those for PHF8 and KDM7, in that two His residues were predicted as active residues. The first one, His335, corresponds to position 254, which is an Asn in both PHF8 and KDM7. The second one, His338, is located three residues N-terminal to the previous one, and corresponds to a His in both PHF8 (His336) and KDM7 (His371).

In JmjD6, Glu231, the residue with third highest score, is highly likely to be involved in substrate docking, as its mutation to Ala was shown to alter substrate hydroxylation in the same study [32]. Demonstration of a lysine-hydroxylation activity for this enzyme does not contradict such a proposal [38]. In JmjD4, Ser232 had the fifth highest score, which makes it a possible candidate for a favorable interaction with the ammonium group of a methylated-lysine substrate. Interestingly, in the JmjN-containing protein JmjD2A, a residue from the JmjN domain, Lys46, was detected as an active-site residue, while in KDM7A, Asp392, which is located on the edge of the JmjC domain and points towards its PHD finger, was also predicted as an active residue. These two results confirmed the importance of the multiple-domain architecture of lysine demethylases for their proper catalytic activity.

Overall, the results underline the relevance of the active-site detection method on the crystallized proteins, since for three of them the three iron ligands are detected, and for the two remaining ones, two iron ligands are detected. This therefore supports the relevance of the other active site residues detected by the method as putative substrate-binding residues, both for the crystallized structures and for the predicted JmjD4 structure.

The present results also underline that information on the domain architecture of a JmjC-containing demethylase can be obtained by considering only the sequence of its JmjC domain (see Discussion), which reveals a concerted within-domain and across-domain evolutionary stress.

DISCUSSION

The present condition-based analysis highlights two important regions in the JmjC sequences, which could be correlated with substrate preference. Four out of the five families have conditions at alignment positions directly N-terminal to position ‘Asn’ (positions 252 and 253), with one condition (253:W) for family IV and two conditions for families I, II and V (Table 1). Residues at these positions should affect the geometry of the substrate-accepting pocket and the surroundings of residue at position 254, which interacts with methylated lysine from substrate when an asparagine [35; 37]. Similarly, all families have conditions at either or both positions 23 and 24, which are directly C-terminal to the second iron ligand (Figure 2, position 22). Note that families I, II, IV and V, which have conditions either at position 253 only or at both 252 and 253, also have conditions at position 23 and 24, thus revealing strong correlations between conditions at these positions.

Relationship between family conditions, fold and substrate specificity for JmjC-demethylases

Different substrate selectivities have been evidenced for JmjC-domain containing demethylases. The different selectivities can be confronted with the sequence-based classification presented here (Table 1 and Figure 2). For family I, discrepancies have been observed between PHF2 on the one hand and PHF8 and KDM7 on the other hand. A recent functional study showed that the nature of the third iron ligand was not critical for this activity [33]. The present study suggests that positions others than the iron ligand could be more critical at differentiating these three proteins, e.g. positions 23 (Glu+1) and 254. Position 23 is indeed involved in condition 23:F from family I (Table 1 and Table 3), which distinguishes PHF2 from PHF8 and KDM7: while the latter two fulfill this condition, PHF2 has a Ser at position 23, i.e. a small, polar residue, instead of a bulky aromatic residue. The consideration of a mutant structure for PHF2 further highlighted the potential role of this position at altering substrate interaction for PHF2 (Supplementary figure S4), together with position 254, as further supported by the active site calculations (Figure 4).

Preliminary structural results obtained here on family II provide some insights into the possible differences this family would have with respect to the families with characterized protein-histone complexes. Both conditions 23:V and 252:N are likely to have a strong influence on substrate recognition in enzymes from this family, due to the spatial location of residues at these positions in family I and III. For family II, it is also possible that residues N-terminal to His¹ be involved in the recognition of a substrate, such as residues Phe222 and Tyr224 from JmjD4, which are conserved across species. To further assert this proposal, residues N-terminal to the core of the JmjC domain should be included in the analysis, when available in published sequences.

Family III includes JmjD2A (KDM4A), which fulfills five of its conditions (note that this protein is absent from the dataset due to the filtering on sequence homology). JmjD2A has mono-demethylase activity on trimethylated HIS3-Lys36 and HIS3-Lys9, as has been observed on HeLa cells [39] and on its drosophila homologues [40]. Moreover, KDM4D was shown to demethylate HIS1.4-Lys26 when di- or tri-methylated [41]. These results suggest a preference for the Ala-Arg-Lys-Ser motif for demethylases from this family, since this sequence is present around both HIS3-Lys9 and HIS1.4-Lys26.

In family IV, demethylation of HIS3-Lys27Me₃ and HIS3-Lys27Me₂ has been described for both KDM6A (UTX) and KDM6B [21; 42; 43]. Considering the conditions that define this family suggests the existence of specific structural features among proteins from this family, with possibly an increased role for hydrophobic interactions. Multiple conditions indeed involve Trp residues at positions where other families have either smaller or less hydrophobic residues (conditions 172:W, 233:W and 253:W; Table 1, Figure 2). The recent determination of a structure for KDM6B supports this proposal, with hydrophobic contacts between Trp at positions 172 and 233 (Trp1435 and Trp1468, respectively), and interactions of Trp at position 253 (Trp1480) with Leu1344 and Val1353 that help position the domain N-terminal to the JmjC domain (PDB entry 2XXZ, Che, Yue, Krojer, Oppermann *et al.*, unpublished).

Considering the different conditions that define family V confirmed the uniqueness of some structural features for this family: most proteins from this family have a Trp at position 231 (His²-1), such as 43 out of the 44 proteins that simultaneously satisfy conditions 252:T and 253:Q, while smaller hydrophobic residues are present at this position in all other families (Figure 2). The meaningfulness of this discrepancy was confirmed by the presence of 231:W among the set of second-percentile conditions for this family (Supplementary table S3). Finally, it should be noted that, in a preliminary docking analysis for JmjD6, peptide HIS3-Arg2Me_{2a}-Lys4Me₂ could be docked to the protein, but with methylated lysine and not arginine presenting a methyl group to the metal ion (Supplementary figure S5). This result is consistent with the proposed lysyl-hydroxylating role of this enzyme [38; 44], as opposed to its earlier attribution as an arginine demethylase [45].

Overall, even though our family classification parallels the observed substrate preferences among JmjC demethylases, an unambiguous attribution of one or multiple preferred substrate(s) for each family is still not possible. This suggests either that a classification involving more families should be used (see Supplementary table S3), or that substrate selectivity is determined by sequence regions not considered in our analysis, e.g. by residues in sequence regions located outside of the catalytic core. It was indeed observed that Gly170 from JmjD2A, which is located five residues N-terminal to the catalytic core of the protein, interacts with the histone substrate [35], as well as Tyr234 and Leu236 from PHF8 [34]. Additional experiments are also needed in order to provide a rationale for the differences observed between *in vivo* and *in vitro* activities, as e.g. for KDM7 [46; 47] and KDM5/Jhd2 [48].

Families of conditions and domain composition

In the present study, the method we developed earlier for the family classification of PHD fingers, chromodomain and acyltransferase domain [27], was applied to JmjC domains. This application further supports the robustness of the method, due to the good classification rate obtained on JmjC domains (Table 2). Interestingly, families of JmjC domains could be related to the domain composition of JmjC domain-containing proteins (Table 2).

Family I includes JmjC proteins that also contain a PHD finger C-terminal to the JmjC domain, such as PHF2, PHF8, KDM7 and JHD1/KDM2. All PHD-JmjC proteins that were assigned to this family satisfied a high number of conditions from this family (9 to 13 conditions, out of a maximum of 14). The assignment of PHF2, PHF8 and KDM7 to this family is in agreement with the classification proposed earlier by Klose and coworkers [24], which relied on whole-protein sequence comparisons, while a more recent classification assigns the JmjC domains of these proteins to different groups [49]. It should also be noted that, when analyzing only the PHD fingers from these proteins, the three of them would be classified in the same PHD subfamily, according to our condition-based classification [27]. With respect to overall architecture, proteins from family II seem to contain a JmjC domain as sole histone-interacting domain.

Trends are also observed for JmjN-containing JmjC-domain proteins, which were mainly assigned to family III. Forty-three proteins are strict members of this family, with e.g. JmjD2A (KDM4A) satisfying 5 conditions from the family, out of a maximum of 10

possible conditions. Still, proteins from family III have diverse domain organizations, with one or two Zinc-Finger or PHD-finger domain C-terminal to the JmjC (e.g. A6ZY60 from *Sacch. cerevisiae*), or an ARID domain between the JmjN and JmjC domains (Jmj2). As for proteins with tetratricopeptide repeats (TPR), they appear to be grouped into family IV, with e.g. KDM6A, KDM6B and C3ZDP8 as members of this family. Overall, these results suggest that information about the domain surrounding the JmjC domain can be obtained solely by the analysis of a few positions from the JmjC domain, such as those involved in the conditions presented in Table 1.

Main conclusions and future directions

We presented a sequence-based classification of histone demethylases JmjC domains. This classification further confirms the relevance of the relative entropy and affinity clustering-based protein subfamily classification algorithm [27], by producing a satisfactory grouping of JmjC domain-containing proteins into five families. These families show that a classification that only uses the sequences of JmjC domains, without considering the surrounding protein domains, as most previous classifications did, is possible. The classification moreover suggests that for most families, the domain architecture of the proteins can be deduced from the a few alignment positions within the JmjC domain. Sequence regions outside of the catalytic core of the JmjC domain are currently under study in order to evidence further important conditions in the JmjC sequences.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The author thanks Pr. Donald Geman for comments both on the methodology and on the manuscript, as well as for funding (NIH-NCRR Grant UL1 RR 025005). David Simcha is also acknowledged for his contribution for the R computer code that transforms PDB files into contact graphs.

REFERENCES

1. Kouzarides T. Chromatin Modifications and Their Function. *Cell*. 2007; 128:693–705. [PubMed: 17320507]
2. Taverna SD, Li H, Ruthenburg AJ, Allis CD, Patel DJ. How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. *Nat Struct Mol Biol*. 2007; 14:1025–1040. [PubMed: 17984965]
3. Robzyk K, Recht J, Osley MA. Rad6-dependent ubiquitination of histone H2B in yeast. *Science*. 2000; 287:501–504. [PubMed: 10642555]
4. Foster ER, Downs JA. Histone H2A phosphorylation in DNA double-strand break repair. *FEBS J*. 2005; 272:3231–3240. [PubMed: 15978030]
5. Nathan D, Ingvarsdottir K, Sterner DE, Bylebyl GR, Dokmanovic M, Dorsey JA, Whelan KA, Krsmanovic M, Lane WS, Meluh PB, Johnson ES, Berger SL. Histone sumoylation is a negative regulator in *Saccharomyces cerevisiae* and shows dynamic interplay with positive-acting histone modifications. *Genes Dev*. 2006; 20:966–976. [PubMed: 16598039]
6. Marmorstein R, Trievel RC. Histone modifying enzymes: Structures, mechanisms, and specificities. *Biochem. Biophys. Acta*. 2009; 1789:58–68. [PubMed: 18722564]
7. Dekker FJ, Haisma HJ. Histone acetyl transferases as emerging drug targets. *Drug Discovery Today*. 2009; 14:942–948. [PubMed: 19577000]

8. Zhu P, Martin E, Mengwasser J, Schlag P, Janssen K-P, Göttlicher M. Induction of HDAC2 expression upon loss of APC in colorectal tumorigenesis. *Cancer cell*. 2004; 5:455–463. [PubMed: 15144953]
9. Patnaik D, Chin HG, Esteve PO, Benner J, Jacobsen SE, Pradhan S. Substrate specificity and kinetic mechanism of mammalian G9a histone H3 methyltransferase. *J Biol Chem*. 2004; 279:53248–53258. [PubMed: 15485804]
10. Secombe J, Li L, Carlos L, Eisenman RN. The Trithorax group protein Lid is a trimethyl histone H3K4 demethylase required for dMyc-induced cell growth. *Genes & Development*. 2007; 21:537–551. [PubMed: 17311883]
11. Agger K, Cloos PA, Christensen J, Pasini D, Rose S, Rappsilber J, Issaeva I, Canaani E, Salcini AE, Helin K. UTX and JMJD3 are histone H3K27 demethylases involved in HOX gene regulation and development. *Nature*. 2007; 449:731–734. [PubMed: 17713478]
12. Qian C, Wang X, Manzur K, Sachchidanand A, Farooq L, Zeng R, Wang M-M, Zhou. Structural Insights of the Specificity and Catalysis of a Viral Histone H3 Lysine 27 Methyltransferase. *Journal of Molecular Biology*. 2006; 359:86–96. [PubMed: 16603186]
13. Couture J-F, Collazo E, Ortiz-Tello PA, Brunzelle JS, Trievel RC. Specificity and mechanism of JMJD2A, a trimethyllysine-specific histone demethylase. *Nat Struct Mol Biol*. 2007; 14:689–695. [PubMed: 17589523]
14. Brown M, Sims R, Gottlieb P, Tucker P. Identification and characterization of Smyd2: a split SET/MYND domain-containing histone H3 lysine 36-specific methyltransferase that interacts with the Sin3 histone deacetylase complex. *Molecular Cancer*. 2006; 5:26. [PubMed: 16805913]
15. Shi Y, Lan F, Matson C, Mulligan P, Whetstine JR, Cole PA, Casero RA, Shi Y. Histone Demethylation Mediated by the Nuclear Amine Oxidase Homolog LSD1. *Cell*. 2004; 119:941–953. [PubMed: 15620353]
16. Tsukada, Y-i; Fang, J.; Erdjument-Bromage, H.; Warren, ME.; Borchers, CH.; Tempst, P.; Zhang, Y. Histone demethylation by a family of JmjC domain-containing proteins. *Nature*. 2006; 439:811–816. [PubMed: 16362057]
17. Huarte M, Lan F, Kim T, Vaughn MW, Zaratiegui M, Martienssen RA, Buratowski S, Shi Y. The Fission Yeast Jmj2 Reverses Histone H3 Lysine 4 Trimethylation. *J Biol Chem*. 2007; 282:21662–21670. [PubMed: 17550896]
18. Tsukada, Y-i; Ishitani, T.; Nakayama, KI. KDM7 is a dual demethylase for histone H3 Lys 9 and Lys 27 and functions in brain development. *Genes & Development*. 2010; 24:432–437. [PubMed: 20194436]
19. Stavropoulos P, Blobel G, Hoelz A. Crystal structure and mechanism of human lysine-specific demethylase-1. *Nat Struct Mol Biol*. 2006; 13:626–632. [PubMed: 16799558]
20. Zhang Y, Yang H, Guo X, Rong N, Song Y, Xu Y, Lan W, Zhang X, Liu M, Cao C. The PHD1 finger of KDM5B recognizes unmodified H3K4 during the demethylation of histone H3K4me2/3 by KDM5B. *Protein Cell*. 2014; 5:837–850. [PubMed: 24952722]
21. Sengoku T, Yokoyama S. Structural basis for histone H3 Lys 27 demethylation by UTX/KDM6A. *Genes Dev*. 2011; 25:2266–2277. [PubMed: 22002947]
22. Johansson C, Tumber A, Che K, Cain P, Nowak R, Gileadi C, Oppermann U. The roles of Jumonji-type oxygenases in human disease. *Epigenomics*. 2014; 6:89–120. [PubMed: 24579949]
23. Zhou X, Ma H. Evolutionary history of histone demethylase families: distinct evolutionary patterns suggest functional divergence. *BMC Evolutionary Biology*. 2008; 8:294. [PubMed: 18950507]
24. Klose RJ, Kallin EM, Zhang Y. JmjC-domain-containing proteins and histone demethylation. *Nat. Rev. Genet*. 2006; 7:715–727. [PubMed: 16983801]
25. Neher E. How frequent are correlated changes in families of protein sequences? *Proc Natl Acad Sci U S A*. 1994; 91:98–102. [PubMed: 8278414]
26. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*. 1999; 286:295–299. [PubMed: 10514373]
27. Slama P, Geman D. Identification of family-determining residues in PHD fingers. *Nucleic Acids Res*. 2010; 39:1666–1679. [PubMed: 21059680]

28. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, et al. The Pfam protein families database. *Nucleic Acids Res.* 2011; 40:D290–D301. [PubMed: 22127870]
29. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics.* 2010; 26:680–682. [PubMed: 20053844]
30. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol.* 1993; 234:779–815. [PubMed: 8254673]
31. Slama P, Filippis I, Lappe M. Detection of protein catalytic residues at high precision using local network properties. *BMC Bioinformatics.* 2008; 9:517. [PubMed: 19055796]
32. Wen H, Li J, Song T, Lu M, Kan PY, Lee MG, Sha B, Shi X. Recognition of histone H3K4 trimethylation by the plant homeodomain of PHF2 modulates histone demethylation. *J Biol Chem.* 2010; 285:9322–9326. [PubMed: 20129925]
33. Horton JR, Upadhyay AK, Hashimoto H, Zhang X, Cheng X. Structural basis for human PHF2 Jumonji domain interaction with metal ions. *J Mol Biol.* 2011; 406:1–8. [PubMed: 21167174]
34. Horton JR, Upadhyay AK, Qi HH, Zhang X, Shi Y, Cheng X. Enzymatic and structural insights for substrate specificity of a family of jumonji histone lysine demethylases. *Nat Struct Mol Biol.* 2010; 17:38–43. [PubMed: 20023638]
35. Chen Z, Zang J, Kappler J, Hong X, Crawford F, Wang Q, Lan F, Jiang C, Whetstone J, Dai S, Hansen K, Shi Y, et al. Structural basis of the recognition of a methylated histone tail by JMJD2A. *Proc Natl Acad Sci U S A.* 2007; 104:10818–10823. [PubMed: 17567753]
36. Yang Y, Hu L, Wang P, Hou H, Lin Y, Liu Y, Li Z, Gong R, Feng X, Zhou L, Zhang W, Dong Y, et al. Structural insights into a dual-specificity histone demethylase ceKDM7A from *Caenorhabditis elegans*. *Cell Res.* 2010; 20:886–898. [PubMed: 20567261]
37. Mantri M, Krojer T, Bagge EA, Webby CA, Butler DS, Kochan G, Kavanagh KL, Oppermann U, McDonough MA, Schofield CJ. Crystal Structure of the 2-Oxoglutarate- and Fe(II)-Dependent Lysyl Hydroxylase JMJD6. *J Mol Biol.* 2010
38. Webby CJ, Wolf A, Gromak N, Dreger M, Kramer H, Kessler B, Nielsen ML, Schmitz C, Butler DS, Yates JR 3rd, Delahunty CM, Hahn P, et al. Jmjd6 catalyses lysyl-hydroxylation of U2AF65, a protein associated with RNA splicing. *Science.* 2009; 325:90–93. [PubMed: 19574390]
39. Whetstone JR, Nottke A, Lan F, Huarte M, Smolikov S, Chen Z, Spooner E, Li E, Zhang G, Colaiacovo M, Shi Y. Reversal of Histone Lysine Trimethylation by the JMJD2 Family of Histone Demethylases. *Cell.* 2006; 125:467–481. [PubMed: 16603238]
40. Lloret-Llinares M, Carre C, Vaquero A, de Olano N, Azorin F. Characterization of *Drosophila melanogaster* JmjC+N histone demethylases. *Nucleic Acids Res.* 2008; 36:2852–2863. [PubMed: 18375980]
41. Trojer P, Zhang J, Yonezawa M, Schmidt A, Zheng H, Jenuwein T, Reinberg D. Dynamic Histone H1 Isoform 4 Methylation and Demethylation by Histone Lysine Methyltransferase G9a/KMT1C and the Jumonji Domain-containing JMJD2/KDM4 Proteins. *J Biol Chem.* 2009; 284:8395–8405. [PubMed: 19144645]
42. Hong S, Cho YW, Yu LR, Yu H, Veenstra TD, Ge K. Identification of JmjC domain-containing UTX and JMJD3 as histone H3 lysine 27 demethylases. *Proc Natl Acad Sci U S A.* 2007; 104:18439–18444. [PubMed: 18003914]
43. Lan F, Bayliss PE, Rinn JL, Whetstone JR, Wang JK, Chen S, Iwase S, Alpatov R, Issaeva I, Canaani E, Roberts TM, Chang HY, et al. A histone H3 lysine 27 demethylase regulates animal posterior development. *Nature.* 2007; 449:689–694. [PubMed: 17851529]
44. Unoki M, Masuda A, Dohmae N, Arita K, Yoshimatsu M, Iwai Y, Fukui Y, Ueda K, Hamamoto R, Shirakawa M, Sasaki H, Nakamura Y. Lysyl 5-hydroxylation, a novel histone modification, by Jumonji domain containing 6 (JMJD6). *J Biol Chem.* 2013; 288:6053–6062. [PubMed: 23303181]
45. Chang B, Chen Y, Zhao Y, Bruick RK. JMJD6 is a histone arginine demethylase. *Science.* 2007; 318:444–447. [PubMed: 17947579]
46. Kleine-Kohlbrecher D, Christensen J, Vandamme J, Abarrategui I, Bak M, Tommerup N, Shi X, Gozani O, Rappsilber J, Salcini AE, Helin K. A functional link between the histone demethylase PHF8 and the transcription factor ZNF711 in X-linked mental retardation. *Mol Cell.* 2010; 38:165–178. [PubMed: 20346720]

47. Lin H, Wang Y, Wang Y, Tian F, Pu P, Yu Y, Mao H, Yang Y, Wang P, Hu L, Lin Y, Liu Y, et al. Coordinated regulation of active and repressive histone methylations by a dual-specificity histone demethylase ceKDM7A from *Caenorhabditis elegans*. *Cell Res.* 2010; 20:899–907. [PubMed: 20567262]
48. Huang F, Chandrasekharan MB, Chen YC, Bhaskara S, Hiebert SW, Sun ZW. The JmjN domain of Jhd2 is important for its protein stability, and the plant homeodomain (PHD) finger mediates its chromatin association independent of H3K4 methylation. *J Biol Chem.* 2010; 285:24548–24561. [PubMed: 20538609]
49. Pedersen MT, Helin K. Histone demethylases in development and disease. *Trends Cell Biol.* 2010; 20:662–671. [PubMed: 20863703]

Figure 2.

Sample alignment for JmjC domain sequences. Boxes represent the iron ligands and highlighting the family conditions (Table 1) with family I as gray, family II as gray blue, family III as peach, family IV as blue and family V as yellow. Most central condition from each family of conditions is highlighted in bold. Protein usual names, or Uniprot names when not available, are given, with organisms indicated with the following suffixes: AT, *Arabidopsis thaliana*, CB, *Caenorhabditis briggsae*, Ch, chicken, CL, *Candida lipolitica*, CE, *Caenorhabditis elegans*, DH, *Debaryomyces hansenii*, Hy, *Hydra (Hydra vulgaris)*, M,

mouse, R, rat and SP, *Schizosaccharomyces pombe*, with human as default. Though not all alignment positions are shown for readability, all positions at which a first-percentile condition originates are displayed. Numbering is indicated when discontinuous.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 3.

Comparison of an amino-acid identity-based clustering with our family assignment for JmjC proteins. The tree was generated by comparing sequences of JmjC domains using BLOSUM62 average distances. Roman numbers above the tree leaves correspond to family assignment (Table 1). All proteins from Figure 1 are shown, in addition to KDM7A from *C. elegans*.

Figure 4.

Proposed active site residues for PHF2. The residues predicted as part of an active or functional site in PHF2 according to a network-based algorithm are highlighted with carbon atoms in yellow, except for the detected iron ligands which have their C atoms in green. The cofactor analogue oxalylglycine is shown with C atoms in purple. The iron-binding Tyr residue is shown with C atoms in cyan. Interaction of Lys259 with the cofactor analogue is shown as yellow dots, that of His335 (C) with Ala255 as grey dots. Interaction of His338 with Ser256 (O at 2.93 from N₂) is not displayed, for readability.

Table 1

Families of first-percentile conditions obtained on the JmjC domains.

I	II	III	IV	V
231:I	44:S	200:N	59:G	252:T
241:T	23:V	97:Q	70:P	23:P
252:G	51:N	217:H	253:W	253:Q
17:D	4:F	230:G	4:Q	6:I
253:G	255:W	82:Q	23:N	
23:F	253:H	228:Y	8:K	
24:G	252:N	21:L	48:N	
84:F	24:F	10:W	233:W	
242:P		11:K	172:W	
8:S		24:Q	13:R	
250:V		24:M	76:D	
251:F		10:C	230:T	
13:Y			230:C	
123:W			4:M	
			27:L	

The first row represents the centroid of the cluster of conditions, as obtained by affinity clustering (see METHODS). Conditions are then ranked by decreasing similarity to centroid. Families were numbered by increasing 'core' size (average dissimilarity to the centroid for the three first conditions closest to centroid).

Table 2

Family assignment of JmjC domains and family features.

Family	Assigned proteins	Strict members	Sample members	Typical architecture
I	56	40	JHD1 KDM7 PHF8	PHD-JmjC
II	37	26	Jmj1 JmjD4	JmjC
III	74	69	KDM4A KDM4D	JmjN-JmjC-PHD/ZnF
IV	83	58	KDM6A KDM6B	TPR-JmjC
V	54	41	JmjD6 PTDSR	JmjC
Total	304	234		

Table 3

Structural roles for 'condition' residues.

Family	Condition	Observed role (actual residue)
I	8 :S	docks cosubstrate (Ser238 PHF8)
I	17 :D	provides hydrophilic environment for Arg8
I	241 :T	docks cosubstrate (Thr323 PHF8)
I	252 :G, 253 :G	enlarges cavity for Lys9Me2 (PHF8)
II	252 :N	docks substrate/cofactor (JmjD4, model)
III	10 :W	hydrophobic interaction with Leu121 from KDM4D, positions helix 117-128 (Trp185 KDM4D)
III	11 :K	interacts with carbonyl from Asn128 of KDM4D, positions helix 117-128 (Lys186 KDM4D)
IV	4 :Q	binds main-chain of Lys27Me ₃ , N _{e2} interacts with main-chain carbonyl (Q1377 KDM6B) in hydrophobic interaction with Me from Lys27Me ₃
IV	8 :K	docks cofactor (K1381 KDM6B)
IV	48 :N, 59 :G, 70 :P, 76 :D	stabilize kink between two β -sheets that surround active site (70 : Pro1405 KDM6B)
IV	13 :R, 172 :W	interactions positions active site β -sheet (Arg1386-Trp1438 KDM6B)
V	252 :T	docks cofactor (Thr285 JmjD6)
V	253 :Q	hydrogen-bonds Tyr150 from JmjD6, positions α -helix 142–150 (Gln286 JmjD6)

Interactions were observed in existing structures for families I, III, IV. For family V cofactor was incorporated into the active site from homologous structures. For family II a model for Jmd4 was used (this study).

Table 4

Detection of functional sites within JmjC domains.

PHF8	PHF2	JmjD2A
Asp249 [*]	Asp251 [*]	Lys206
His319 [*]	His335	His188 [*]
His336	His338	His276 [*]
His247 [*]	His249 [*]	Glu190 [*]
	Lys266	Asp193
		Lys46
KDM7A	JmjD4	JmjD6
Asp497 [*]	Asp237 [*]	Asp189 [*]
His567 [*]	<u>His235</u> [*]	Lys204
His495 [*]	<u>Lys253</u>	Glu231
Asp392	<u>His315</u> [*]	Asp158
His584	<u>Ser232</u>	His273 [*]
Glu426		
Asp493		
Glu480		
Lys512		
His620		

Residues terminated with * indicate an iron ligand. Underlined residues are those that possessed a score inferior to the actual *MDevI* value for the protein. See METHODS for the PDB structure used for each protein. Position 77 corresponds respectively to Lys512 in KDM7A, Lys266 in PHF2, Lys253 in JmjD4, Lys206 in JmjD2A and Lys204 in JmjD6.