Structural insights into the regulation and the recognition of histone marks by the SET domain of NSD1

Masayo Morishita, Eric di Luccio*

School of Applied Biosciences, Kyungpook National University, Daegu 702-701, South Korea

ARTICLE INFO

Article history:
Received 13 July 2011
Available online 22 July 2011

Keywords:
Epigenetic therapy of cancer
Histone-lysine methyltransferase
NSD1
NSD2/MMSET/WHSC1
NSD3/WHSC1L1
HMTase inhibitors

ABSTRACT

The development of epigenetic therapies fuels cancer hope. DNA-methylation inhibitors, histone-deacetylase and histone-methyltransferase (HMTase) inhibitors are being developed as the utilization of epigenetic targets is emerging as an effective and valuable approach to chemotherapy as well as chenoprevention of cancer. The nuclear receptor binding SET domain (NSD) protein is a family of three HMTases, NSD1, NSD2/MMSET/WHSC1, and NSD3/WHSC1L1 that are critical in maintaining the chromatin integrity. A growing number of studies have reported alterations or amplifications of NSD1, NSD2, or NSD3 in numerous carcinogenic events. Reducing NSDs activity through specific lysine-HMTase inhibitors appears promising to help suppressing cancer growth. However, little is known about the NSD pathways and our understanding of the histone lysine-HMTase mechanism is partial. To shed some light on both the recognition and the regulation of epigenetic marks by the SET domain of the NSD family, we investigate the structural mechanisms of the docking of the histone-H4 tail on the SET domain of NSD1. Our finding exposes a key regulatory and recognition mechanism driven by the flexibility of a loop at the interface of the SET and postSET region. Finally, we prospect the special value of this regulatory region for developing specific and selective NSD inhibitors for the epigenetic therapy of cancers.

© 2011 Elsevier Inc. All rights reserved.

1. Introduction

1.1. Transcriptional regulation by covalent histone modifications

Histones are the stage of diverse post-translational modifications that ultimately regulates the gene transcription. Lysine methylation is one prominent feature of the post-translational histone modifications in the regulation of chromatin structure and function. Lysine-histone methyltransferases (HMTases) target specific histone residues on H3 and H4, and can transfer one, two or three methyl groups on specific lysines on the histone tails [1]. Lysine or arginine methylation, or any of the other histone modifications, can have both activating and repressive functions on transcription. All the covalent histone modifications contribute to regulating the diverse activities associated with the chromatin and may be referred as a language of covalent histone modifications or histone code [2].

Both genetic and epigenetic alterations of transcription factors are key features in carcinogenesis onset with aberrant gene functions and changes in gene expression levels [3]. Many such as NSD1, NSD2/MMSET/WHSC1, NSD3/WHSC1L1, Suv39h, Ezh2, Mll, Riz, and others HMTases are implicated in numerous tumor developments and are generally found over-expressed in tumorous tissues [4]. Usually, reverting HMTase levels or inhibiting HMTase activity appears to be promising in helping suppressing cancer growth [3,5]. Therefore, the development of HMTase inhibitors is emerging as an effective strategy in the epigenetic therapy of cancers.

1.2. The NSD family of HMTase

The NSD family consists of three NSD proteins that are encoded at the loci 5q35 (NSD1), 4p16 (NSD2), and 8p12 (NSD3). Several transcripts have been identified (NSD1; 17, NSD2; 29, NSD3; 5) whose exact biological functions remain unclear. NSD1 histone-lysine N-methyltransferase (lysine-HMTase) specifically mediates the methyl transfer onto a H3 lysine-36 and H4 lysine-20 (H3K36 and H4K20), NSD1 is a 2696-AA protein that contains four zinc finger domains, two PWPF domains, and a SET domain responsible for the HMTase activity (Fig. 1A). NSD2 (lysine-HMTase H3K4 and H4K20) and NSD3 (lysine-HMTase H3K36) share the same functional domains with NSD1, but on a shorter sequence of 1365-AA (NSD2) and 1437-AA (NSD3), respectively (Fig. 1A). NSD1 functions in both the retinoic acid receptor (RAR) activation

Abbreviations: NSD, nuclear receptor binding SET domain; HMTase, histone-lysine methyltransferase; AdoMET, S-adenosylmethionine; MD, molecular dynamics; RMSD, root mean square deviation.

* Corresponding author. Fax: +82 53 950 6750.
E-mail address: eric.diluccio@gmail.com (E. di Luccio).

0006-291X/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
and the lysine degradation pathways [6]. Both NSD2 and NSD3 operate solely in the lysine degradation pathway [6]. However, despite sharing a strong similarity with NSD1, NSD2 and NSD3 are likely to operate in distinct functional pathways as both NSD2 and NSD3 do not make substantial contribution to overgrowth phenotypes observed with NSD1 gene alterations [7]. It remains unclear whether any of the NSD family members can substitute each other in functional pathways.

1.3. Cancers associated with NSD1/NSD2/NSD3

Other and we have recently reviewed the role of NSD proteins as oncogenes [8,9]. A growing number of studies link the NSD proteins to a variety of cancers [2–4,7–10,14–17,19–28]. NSD1 is associated with acute myeloid leukemia, lung cancer, neuroblastomas and glioblastomas [4,11–17]. NSD2 has been found associated with the prostate cancer and multiple myeloma [4,18–21]. Furthermore, the amplification of either NSD1 or NSD2 trigger the cellular transformation, initiating carcinogenesis events [4,11–13,15–17]. Increased NSD2 activity was reported in the tumor proliferation in glioblastoma multiforms [22]. Overexpression of NSD2 in myeloma cells leads to aberrantly high global levels of H3K36 di-methylation, accompanied by a decrease in H3K27 methylation. In myeloma cells, NSD2 contributes to disrupt the chromatin structure and function contributing to the cellular transformation [23]. In addition, NSD2 is found overexpressed in 15 different cancers and is associated with tumor aggressiveness or prognosis in most types of cancers [10,20]. NSD3 is found amplified in breast cancer cell lines and

Fig. 1. The NSD family (A) Schematic of the primary structure of NSD1, 2 and 3: PWWP domain; PHD zinc fingers domain; SET histone methyl transferase (HMTase) with the preset and postSET domains. The regulatory loop closing onto the histone-binding site is indicated in red. (B) Multiple sequence alignment of the preSET, SET and postSET domains of NSD1, NSD2, NSD3, MLL1, G9a EHMT2 H3-K9-HMTase, SETD8 H4-K20 HMTase, SETD7 H3-K4-HMTase. Boxed in blue are the dissimilar residues amongst the SET domains of NSD1, NSD2 and NSD3. Dashed boxed in blue indicated the residues involved in the binding of the 1-lysine–histone ligand. Boxed in red are the conserved residues responsible for the binding specificity of both 1-lysine-histone and AdoMet. The residues participating in the regulatory loop are colored in purple. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
primary breast carcinomas [12,24]. Moreover, NSD3 is involved in lung cancer and the acute myeloid leukemia where NSD3 is fused with NUP98, similarly as NSD1 [4,12,18–21].

1.4. The NSD pathway as a target in the epigenetic therapy of cancers

The development of epigenetic therapies is raising hope in cancer therapies. DNA-methylation inhibitors, histone-deacetylase and histone-methyltransferase inhibitors are being developed as the utilization of epigenetic targets is becoming an effective and valuable approach to chemotherapy and the chemoprevention of cancer. The level of histone acetylation is regulated by histone acetyltransferases (HATs) and histone deacetylases (HDACs). Inhibitors of HDACs are able to modulate the expression of specific genes and HDAC inhibitors are being used in cancer therapy. Moreover, HMTases are becoming valuable therapeutic target since deregulated HMTase plays key roles in carcinogenesis onset [25].

An increasing number of studies link deregulated NSD proteins to several cancers and the NSD pathways are focusing attention for the design of specific and selective HMTase inhibitors [8]. However, little is known about the NSD pathways and very few lead compounds have been described to selectively inhibit HMTase enzymes. To the best of our knowledge, none of HMTase inhibitors have been reported to selectively inhibit the NSD proteins. This is mostly due to the lack of structural studies on HMTases in general and the NSD family in particular, that would support the effective design of selective and specific inhibitors.

1.5. Substrate specificity and backbone contributions in the postSET region

The sequences of the SET domain of NSD family is phylogenetically distinct to other known lysine-HMTases (Fig. 1B) [8]. However, the overall scaffold of the SET domain of NSD1 is conserved and both NSD2 and NSD3 are likely to share the same structural feature as NSD1 due to the high sequence identity amongst the NSD SET family (Fig. 1B).

The canonical SET domain of methyltransferases is subdivided into the preSET, SET and postSET regions (see below). The structural mobility within the SET and postSET regions is critical for the function of H3K9 methyltransferases and for the recognition of the substrate [26]. In H3K9-HMTases, the postSET region exists in three conformation state: (1) a flexible/ordered state when no peptide of co-factor is bound; (2) An intermediate state when only the co-factor is bound; and (3) a structured conformation with the whole mobile postSET region closes onto the peptide substrate [26]. In the H3K4-HMTase MLL1, the SET domain undergoes a similar open-closing mechanism [27].

NSD1 differs from other known HMTases. Analysis of the recent crystal-structure of the peptide-less SET domain of NSD1 revealed that in absence of ligand, the histone-binding site is occluded preventing any access to the catalytic groove [28]. Therefore, we hypothesized that the SET domain of NSD1 has specific mechanisms to recognize histone marks and to grant access to the histone-binding site.

Little is known about the NSD family. As a first step toward understanding the regulation and the recognition of histone marks by the SET domain of the NSD family, we use computational methods to investigate the structural mechanisms happening in the SET domain during the binding of the H4-histone tail, exploiting the recent crystal structure of the peptide-less SET domain of NSD1 [28]. We prospect the special value of our findings for developing specific and selective NSD-HMTase inhibitors.

2. Material and methods

2.1. Molecular modeling, docking and energy minimization

A structural overlay between SET-NSD1 (PDB: 3OOI) and the crystal structure of the SET8 bound to histone H4 peptide (PDB: 3F9Y) was used to manually place the H4-peptide residues 16–22 (H4-peptide16–22) into the histone-tail binding site of NSD1 and the cofactor S-adenosylmethionine (AdoMet) into its specific binding pocket, using the SSM algorithm in COOT [29]. The complex H4-peptide16–22–AdoMet–NSD1–SET was subjected to 10,000 steps (5 ps) of restrained torsion molecular dynamics (MD) to ideally position the complex into the binding site groove of the NSD-SET domain (CNS V1.2) [30]. The harmonic restraints was K = 10 kcal/mol Å². After the MD simulations, the complex H4-peptide16–22–AdoMet–NSD1–SET was energy minimized with spatial restraints (CNS V1.2) and manually checked for ideal stereochemistry with COOT [29,30]. The Dundee PRODRG Server was used to generate the CNS topology and parameters files for the cofactor AdoMet [31].

2.2. Long molecular dynamics simulations in a water box

The complex H4-peptide16–22–AdoMet–NSD1–SET was fully solvated with a water box using VMD 1.8.7 [32]. The molecular system used for MD contains 53,227 atoms, of which 49,428 are from the water molecules. The solvated complex H4-peptide16–22–AdoMet–NSD1–SET was further subjected to 900,000 steps (900 ns) of molecular dynamics with NAMD v2.7 using the CHARMM forcefield [33,34]. A constant temperature was maintained by using a Langevin damping at 300 K, and the pressure was held constant at 1 atmosphere with a Langevin piston. The system was equilibrated with 1000 steps of conjugate gradient minimization before MD runs. After the MD simulations, the solvated complex H4-peptide16–22–AdoMet–NSD1–SET was manually checked for ideal stereochemistry with COOT [29,30]. The MD trajectory was analyzed with VMD 1.8.7 [32].

2.3. Model analysis and validation

The details of the interactions between NSD-SET and H4-peptide16–22 were analyzed using Ligplot [35]. The confirmation of the interaction maps was done manually using visual inspections of the models in COOT [29].

3. Results and discussion

The canonical structure of the SET domain is composed of three groups of canonical β-sheets closely neighboring a conserved α-helix defining a cleft for the binding to the histone-tail ligand. The cofactor AdoMet and substrate bind at two adjacent sites to the SET domain (Fig. 2A). The histone-tail ligand binds into a groove formed by both the SET and postSET domains (Fig. 2A–C). The cofactor AdoMet binds into a distinct pocket located on the other side of the SET domain and acts as a methyl group donor. Both AdoMet and the L-lysine-histone are connected through a narrow tunnel where the methyl group is channeled (Fig. 2A). In histone-lysine methyltransferase, the lysine substrate extends at the mouth of the tunnel.

In the crystal structure of the peptide-less NSD-SET domain, the postSET domain loop is extended on top of the histone binding site, which sterically prevents the binding of either H3K36 or H4K20 substrates, unlike in both H3K9 and H3K4-HMTases [26–28]. This particular observation led us to hypothesize that the postSET loop
region of NSD1 might work with regulation purposes, blocking the binding site in the absence of ligand. To shed some light on the function of postSET loop region on the regulation of NSD1, we used long molecular dynamics simulations in a water box at pH7 to accurately model the docking between H4-peptide16–22 and NSD1×SET. The H4-peptide tail for residues 16–22 was manually placed underneath the postSET loop of the peptide-less crystal structure, deliberately bringing the complex NSD1×SET–H4-peptide16–22 in a sterically unstable state. Energy minimizations and long molecular dynamics simulations relieved the steric clashes and force the SET domain to accommodate the substrate (Figs. 2–4 and S1). The MD simulations were long enough to let the complex stabilizes into a stable conformation with little RMSD oscillations for the postSET region (Fig. S1). Interestingly, the only structural modifications are focused on a loop at the interface of the SET and postSET regions (residues 2063–2082) which rotated ~45° and translated ~7 Å at the tip, which is a rather large motion (Figs. 1A and 2B and C). Besides the flexible loop 2063–2082, the backbone of NSD1×SET did not undergo significant structural modifications (overall cr–RMSD < 0.6 Å).

The H4-peptide16–22 fits tightly in the histone binding region with the side chain of lysine 20 extended inside the channel toward the cofactor binding region, as expected. Interestingly, this loop comprised from residues 2063 to 2082 acts both as a seat belt to firmly anchor the ligand in the binding domain (Figs. 2 and 3). One other function of this particular loop is to perhaps contribute to the discrimination and specificity of the substrate by acting as a regulatory loop. In the absence of peptide ligand, the histone-binding site is occluded and the access to the AdoMet channel is obstructed (Fig. 3). The peptide-binding site becomes accessible only after the rotation and displacement of the regulatory loop (Fig. 3). It exposes a largely negatively charged surface area suitable for the docking of the positively charged H3 and H4 tail. The

H4-peptide16–22 is firmly stabilized by a network of interactions to allow the lysine 20 to extend toward the cofactor binding site in order to react (Fig. 4). Lys20–H4-peptide16–22 is anchored by two hydrogen bonds with backbone oxygens of Arg2018 and Met2020 (Fig. 4). This double hydrogen bond is conserved amongst lysine-HMTases. In peptide bound structures of H3K9-HMTases (GLP and Dim-5), the Lys9 is stabilized by a pair of analogous H-bonds [26]. A similar conserved double hydrogen-bond is observed in the structures of H4K20-SET8, H3K4-SETD7 and H3K27-vSET [26]. However, for other lysine-HMTases, the contributions of the backbone and side-chains to the binding of the peptide–substrate is significantly less than for NSD1×SET. The regulatory loop closing on top of the peptide contribute greatly to the stability of the complex with an extended network of interactions with Tyr2059, Asn2060, Leu2061, Cys2063, Asn2066 and Thr2069 (Figs. 3 and 4).

The catalytic mechanism of methyl transfer from AdoMet to the ε-amino of the target lysine by lysine-HMTase proceeds by a linear SN2 nucleophilic attack [36]. Two and three iterations provide dimethylated and trimethylated lysine, respectively. All lysine-HMTase appear to share a similar enzymatic mechanism. However, the recognition of the substrate differs amongst lysine-HMTase as evidenced by the sequence heterogeneity of the SET domain.

HMTase inhibitors are scarce but one of the most efficient strategy seems to target the ε-lysine-histone or ε-arginine-histone binding pocket over the cofactor binding site [5]. The inhibitors BIX-01294 and BIX-01338 have been proven effective on G9a with an IC50 of 3 and 5 μM, respectively [5]. In addition, chaetocin inhibits Su(var)³–9 with an IC50 of 0.8 μM [5]. Unlike the peptide binding area, the cofactor binding area appears to be well conserved across lysine-HMTases. Therefore, AdoMet-derivatives are likely to be less specific than the inhibitors targeting the histone-tail binding site.

---

**Fig. 2.** The conformations of the SET-domain of NSD1. (A) Overall structure of the SET domain. The structure of the SET domain is composed of three groups of canonical β-sheets arranged in a triangular fashion with a group of two β-sheets closely neighboring a conserved α-helix defining a cleft for the binding of the lysine-histone ligand. The cofactor AdoMet binds into a cavity adjacent to the SET domain connected through a channel. The flexible regulatory loop is colored in red. (B) Motion of the regulatory loop. Structural overlay between the peptide-less crystal structure (colored in blue) and the model of NSD1 SET bound with H4-peptide16–22 (colored in blue). The loop motion is indicated, from the peptide-less conformation (colored in orange) to conformation with H4-peptide16–22 bound (colored in red). (C) Model of NSD1-SET bound with H4-peptide16–22 after molecular dynamic simulations. Residues 16–22 of H4K20 are displayed in sticks. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
The design of selective HMTase inhibitors relies on exploiting dissimilarities in the histone-tail binding pocket. SET-containing enzymes are structurally diverse regarding the postSET region. The postSET domains of both G9a and GLP have an $\alpha$-helix that contributes to peptide binding where other HMTases have a loop [37]. In PRMT2, the postSET domain is absent. However, the postSET domain is partially folded in G9a and SUV39H2 bound with only the cofactor. Interestingly, Dot1 is a H3K76 methyltransferase with no SET domain [37]. Therefore, Dot1 is a H3K76 methyltransferase with a regulatory loop restricting the access to the ligand binding area. In the presence of ligand, a network of residues stabilizes the H4-peptide tail on the binding site. Therefore, it is likely that the regulatory loop participates in the discrimination of substrate of NSD1-SET and this specificity should be exploited for the design of inhibitors.

The opening mechanism of the SET domain of NSD1 remains unclear. What is the driving force for displacing the regulatory loop? Our data suggest the binding of the histone-peptide itself is not entirely responsible for the regulatory loop movement. The binding of nucleosomal DNA is perhaps the allosteric effector that triggers the opening of the SET domain. The SET domains of NSD2 and NSD3 are highly related to NSD1 and it is possible all share a similar opening mechanism.

**Acknowledgments**

This study was supported by the Basic Science Research Program (2010-010837) through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology. The Kyungpook National University contributed to support the research presented in this paper. Dr. Phillip B. Carpenter is acknowledged for interesting discussions about the NSD1 pathway.

**Appendix A. Supplementary data**
