ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 18(12): 1146-1153 (2022)

BIOINFORMATION
Discovery at the interface of physical and biological sciences

OPEN ACCESS GOLD

BIOMEDICAL INFORMATICS

www.bioinformation.net
Volume 18(12)

Research Article

Received November 1, 2022; Revised December 20, 2022; Accepted December 31, 2022, Published December 31, 2022
DOI: 10.6026/973206300181143

Declaration on Publication Ethics:
The author's state that they adhere with COPE guidelines on publishing ethics as described elsewhere at https://publicationethics.org/. The authors also undertake that they are not associated with any other third party (governmental or non-governmental agencies) linking with any form of unethical issues connecting to this publication. The authors also declare that they are not withholding any information that is misleading to the publisher in regard to this article.

Declaration on official E-mail:
The corresponding author declares that lifetime official e-mail from their institution is not available for all authors

Comments from readers:
Articles published in BIOINFORMATION are open for relevant post publication comments and criticisms, which will be published immediately linking to the original article without open access charges. Comments should be concise, coherent and critical in less than 1000 words.

Edited by P Kangueane
Citation: Dey & Prasad, Bioinformation 18(12): 1146-1153 (2022)

Structure based functional annotation of a MYND-less lysine methyl transferase in *Candida albicans*

Joydeb Dey & Himanshu Kishore Prasad*

Department of Life Science and Bioinformatics, Assam University, Silchar, Assam-788011, India; *Corresponding author

Institutional URL:
https://www.aus.ac.in/

Author contacts:
Joydeb Dey – E-mail: joydeb.dey@aus.ac.in
Himanshu Kishore Prasad – E-mail: himanshu.k.prasad@aus.ac.in

Abstract:
Candida albicans is opportunistic pathogenic yeast that is widely distributed throughout the world and is classified as the most critical fungal pathogen group. *Candida albicans* is a common microbiota of healthy individuals but can cause superficial and invasive infections in immune compromised individuals. Protein Post-translational modifications involving methylation of lysine amino acids stand for a major regulator of eukaryotic transcription, and pathways controlling several cellular processes. SMYD makes up a SET (Su (Var) 3–9, Enhancer-of-zeste and Trithorax) and MYND (Myeloid, Nervy, and DEAF-1) domain containing lysine methyl transferase subfamily that transfers methyl groups from methyl donors onto lysine residues in histones (H3 and H4) and non-histone proteins. The SET domain is the methyl

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 18(12): 1146-1153 (2022)

transferase catalytic domain, while MYND participates in both protein and DNA interactions. Well-studied examples of SMYD proteins are five human and two *Saccharomyces cerevisiae,* constituting examples of histone and non-histone protein lysine methyl transferase members. However, there is limited understanding of SET lysine methyltransferases, including the SMYD subfamily, in the pathogenic fungi *Candida albicans.* Using bioinformatics tools, we characterized the SMYD domain containing proteins in the important pathogen. We report the presence of an atypical SMYD member (CaO19.3863) as a new lysine methyltransferase that can be a target for antifungal therapy.

Keyword: SMYD, non-canonical Lysine methyltransferase, *Candida albicans*, uncharacterized proteins.

Background:

SMYD proteins are a family of lysine methyltransferases that are characterized by a SET and MYND domain. S-Adenosylmethionine (SAM) is a methyl donor that is bound to the SET domain by many lysine N-methyltransferases **[1]**. The MYND domain is a unique domain that is specific to SMYD proteins, and it is responsible for the recognition of the target lysine residue. The structure of the SET and MYND domains of SMYD proteins is highly conserved among different isoforms, and they are arranged in a characteristic fold that is essential for their enzymatic activity. In particular, the SET domain is composed of an alpha/beta barrel that is stabilized by a conserved zinc ion, while the MYND domain is composed of a helical bundle that is stabilized by two conserved cysteine residues **[1,2]**. The enzymatic activity of SMYD proteins is mediated by the formation of a catalytic triad comprising the SET domain, the MYND domain, and a lysine residue on the target protein. The SET domain binds to SAM and transfers the methyl group to the lysine residue, while the MYND domain facilitates the recognition and positioning of the target lysine residue **[1]**. In addition, the MYND domain contains a conserved hydrophobic patch that is thought to interact with the methylated lysine residue. Moreover, the basic residues in MYND domain contribute to DNA binding and the MYND domain of yeast Set5 (Fungal SMYD family) is its mediator in chromatin association and gene repression near telomere **[3]**. The C-terminal domain of different SMYD members is also involved in protein-protein interactions, localization and control methyltransferase activities in some SMYD members **[3]**. The five-membered human SMYD family of protein lysine methyltransferases (SMYD1-5) have established roles in muscle immune, blood, heart, vascular endothelial physiology, host-pathogen interaction, and in multiple cancer patho-physiology **[4,5]**. Their ability to methylate specific lysine residues on histone proteins (K4,36,37 Histones 3 and K5, Histones 4) and hon-histone proteins (HSP90, Rb, P53, MAP3K2) is crucial for regulating gene expression and protein function in cells **[4-6]**. Overexpression or deregulation of the human SMYD enzymes has been linked to the emergence and progression of cancer, making them promising targets in cancer therapy **[5-7]**. SMYD members with structural similarity in *Saccharomyces cerevisiae* (Set5 and Set6) have similar domain arrangement like their human counterparts. The Set5 enzyme is involved in genome stability and gene expression regulation near telomere mediated by its histone H4 are K5, K8 and K12 methylation activities **[3]**. In the related opportunistic yeast *Candia albicans,* the SMYD family of KMT's are relatively uncharacterized, with *Candia albicans* SET6 expression being regulated by the transcription factor Hap43 and conditions of biofilm and catheters **[8]**. In this paper, the uncharacterized SMYD family proteins of this yeast was considered for *in silico* functional and structural characterizations. Using sequence and structure analysis tools, the 3D models were used for identification of residues involved in ligand binding. In this work, we uncover and characterize a novel SMYD member, conserved in the *Candida* clade organism, which has structural conservation with known SMYD members but completely lacks the MYND zinc finger spanning the SET domain. This protein can be a target for biochemical, genetic, experimental analysis or any future antifungal therapy experiments to ascertain the function.

Methodology:

SET domain methyltransferase repertoire of *Candida albicans*

HMMER 3.1b2 hmmsearch was used to identify all SET domain proteins **[9]**. Using the Pfam hmm profile (accession PF00856.31), the reference protein database of *Candida albicans* SC5314 (taxid: 237561) was scanned **[9]**. The hits were further checked for similarity searches using NCBI BLASTP **[10]**.

Functional analysis of SMYD Proteins

UniprotKB was used to download well-studied SMYD domain proteins from Human, Mouse, Arabidopsis, toxoplasma, and yeast. As part of InterProscan **[11]** analysis, these SMYD domain proteins were further functionally annotated by their superfamilies, families, domains, folds, and motifs to identify conserved domains. We also used the HHpred tool for structural annotations using HMM-HMM searches **[12]**. With ProtParam **[13]**, the physicochemical properties were discovered. Polyphobius was **[14]** for homology-supported predictions of transmembrane topology, Yloc and DeepLoc 2.0, tool for detecting subcellular localization **[15,16]**. In addition, MULocDeep, for suborganelle level localization **[17]** and SignalP-6.0 **[18]** to predict signal peptides. DeepTMHMM assigned transmembrane helices **[19]**. PSIPRED workbench utility, MEMPACK predicted transmembrane helix contact, and DISOPRED3 intrinsic disorders **[20]**. DeepCoil and Lupas were used to predict coiled-coil domains **[21]**.

Multiple Sequence Alignment (MSA), Phylogenetic analysis and Conserved Motifs Identification

MSAs were generated with MUSCLE **[22]**, alignments were edited with trimAl in PhyloSuite v1.2.3pre1 **[23]**, substitution model selection and bootstrapped ML phylogenetic tree was constructed using MEGA version 11.0.13 **[24]**. Conserved sequence motifs were generated using the MEME (Multiple Em for Motif Elicitation) server to predict conserved motif patterns in SMYD proteins **[25]**.

Secondary structure and other protein features prediction

The PredictProtein webserver was used for features such as protein-protein and protein-DNA binding sites, disorder, and metal

binding sites **[26]**. PSIPRED 3.2 was employed for secondary structure prediction using neural networks, and CysPRED **[27]** was employed to detect disulfide bonds. **[27]**. Protein interactions were discovered through physical and functional correlations using STRING 11.5 **[28]**. The LambdaPP pipeline **[29]** was used to annotate gene ontology (GO), binding residues, secondary structure, and variant effect scores.

Tertiary structure determination and Binding sites predictions

Colabfold notebook for AlphaFold V2 **[30] [31]** was used to determine the 3D structures of SMYD proteins. Intradomain confidence was derived from predicted LDDT (pLDDT), while domain confidence was derived from a Predicted Aligned Error (PAE) **[31]**. With AlphaFill **[32]**, ligands, metal ions, and cofactors, information was fitted to high-ranking AlphaFold models. FoldSeek was used for comparative structural analysis **[33]**. Molviewer was used for visualizations **[34]**. GalaxyWEB **[35]** docking tool GalaxySite was used for ligand prediction. The Computed Atlas of Surface Topography of Proteins was used for active site determinations **[36]**. LambdaPP and Predictprotein were used to predict catalytic and SAM, polypeptide, and metal binding sites.

Results and discussion:

The SET domain [Su (var) 3-9, zeste enhancer, Trithorax] containing methyltransferases (KMT) of the *Candida albicans* proteome was identified by sensitive hmmsearch using the hmm profile for SET domain. This search strategy gave eight distinct SET domains containing KMTs **(Table 1)**.

Table 1: *Candida albicans* SET domain methyltransferase from HMMSEARCH and their annotation. The Orthologs were detected using BLASTP searches, and domain positions assigned using PROSITE profile matches.

Accession (e-value)	Length	Systematic Name/Identifier / Name	Features	*S. cerevisiae* /Human Hits	PDB Hits	SET Domain
A0A1D8PNN9_CANAL (5.1e-19)	552	C5_03250W_A / orf19.2654/ CaRKM4	Ribosomal protein lysine N-methyltransferase	RKM4 / SETD6	3QXY_A (Human)	24-288
A0A1D8PP54_CANAL (2.8e-16)	433	C5_05150C_A / orf19.4007/ CaRKM2	Protein-lysine N-methyltransferase	RKM2/SETD4	6ICT_A (Human)	33-281
Q5ABG1/SET1_CANAL (3.9e-15)	1040	C1_00960C_A / orf19.6009 / CaSET1	Histone H3-lysine N-methyltransferase.	SET1/SETD1B & A	6VEN_N (Yeast)	898-1015
Q59XV0/SET2_CANAL (1.2e-14)	844	C2_10250C_A / orf19.1755 / Ca SET2	Histone-lysine N-methyltransferase specific to Histone H3 lysine-36.	SET2/SETD2	6NZO_S (Fungi)	142-259
Q59VZ3_CANAL (4.5e-12)	379	C1_02080W_A / orf19.3665 / CaSET6	Protein of unknown function.	SET6/SMYD4	3MEK_A (Human)	24-346
Q5A1M3/SET5_CANAL (9.9e-09)	473	C5_00950C_A / orf19.1972/ CaSET5	SAM dependent methyltransferase.	SET5/SMYD2	5KJK_A (Human)	107-381
A0A1D8PTX3_CANAL (4.1e-08)	579	CR_09310W_A / orf19.7326 / CaRKM1	Protein-lysine N-methyltransferase Mono & di-methylation	RKM1/SETD3	6V62_A (Human)	52-269
A0A1D8PT54_CANAL (7.4e-05)	630	CR_06100C_A / orf19.3863/ CaSMYD	SET domain-containing protein, ORF, Uncharacterized	SET5/SMYD3	5CCM_A (Human)	180-287

The hits were exhaustively searched by blastp based homology against Uniprot, PDB and nonredundant databases. There are five SETs and three SMYD domain proteins in the *Candida albicans* genome that constitutes KMT proteins **(Table 1)**. Comparatively, *Saccharomyces cerevisiae* has only two SMYD domain proteins and five SET proteins **(Table 1)**. In *Saccharomyces cerevisiae*, Set5 & Set6 are known KMTs of the SET and MYND domains (SMYD) sub-family. *Candida albicans* have CaSet6 (orf19.1972) and CaSet6 (orf19.3665). This work identified the unique third hit of *Candida albicans* SMYD KMT (orf19.3836), which has a truncated SET-MYND (SMYD) domain (107 amino acids) and is uncharacterized. The manuscript discusses computational functional and structural characterization of CaSMYD as a novel member of SMYD domain KMT. To identify closely related sequences and estimate its conservation among other SMYD proteins and in budding yeast, the CaSMYD protein sequence was used as a blastp query. The blastp search produced human SMYD3 (14% coverage and 39.13%), mouse SmyD1 (15% coverage and 34.51% identity), human SMYD2 (15% coverage and 35.791% Identity), and AKMT *Toxoplasma gondii* (21% coverage and 19.71% identity).

Table 2: List of Protein Data Bank (PDB) and Uniport hits with CaSMYD blastp

Protein	Organism	Query Coverage (%)	Identity (%)	ID
SMYD3	*Homo sapiens*	14	39.13	5CCM, 3OXL,6YUH, 5XXD, 3QWP, 5CCL, 6ZRB, 5HI7, 3RU0, 5EX0, 3OXF, 3OXG, 5EX3
SMYD3	*Homo sapiens*	14	38.04	3MEK
SmyD1	*Mus musculus*	15	34.51	3N71
SMYD2	*Homo sapiens*	15	35.79	6CBX, 4WUY, 3S7B, 3RIB, 5ARF, 5KJK, 3TG4, 4YND
SmyD2	*Mus musculus*	15	35	3QWV
Set6	*Schizosaccharomyces pombe*	14	35.87	O94256
Set5	*Schizosaccharomyces pombe*	22	22.29	O74467
AKMT	*Toxoplasma gondii*	21	19.71	6FND
SDG41	*Arabidopsis thaliana*	17	27.33	Q3ECY6.1
ASHR1	*Arabidopsis thaliana*	35	24.02	Q7XJS0.2
ASHR2	*Arabidopsis thaliana*	15	23.58	Q9ZUM9.3

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 18(12): 1146-1153 (2022)

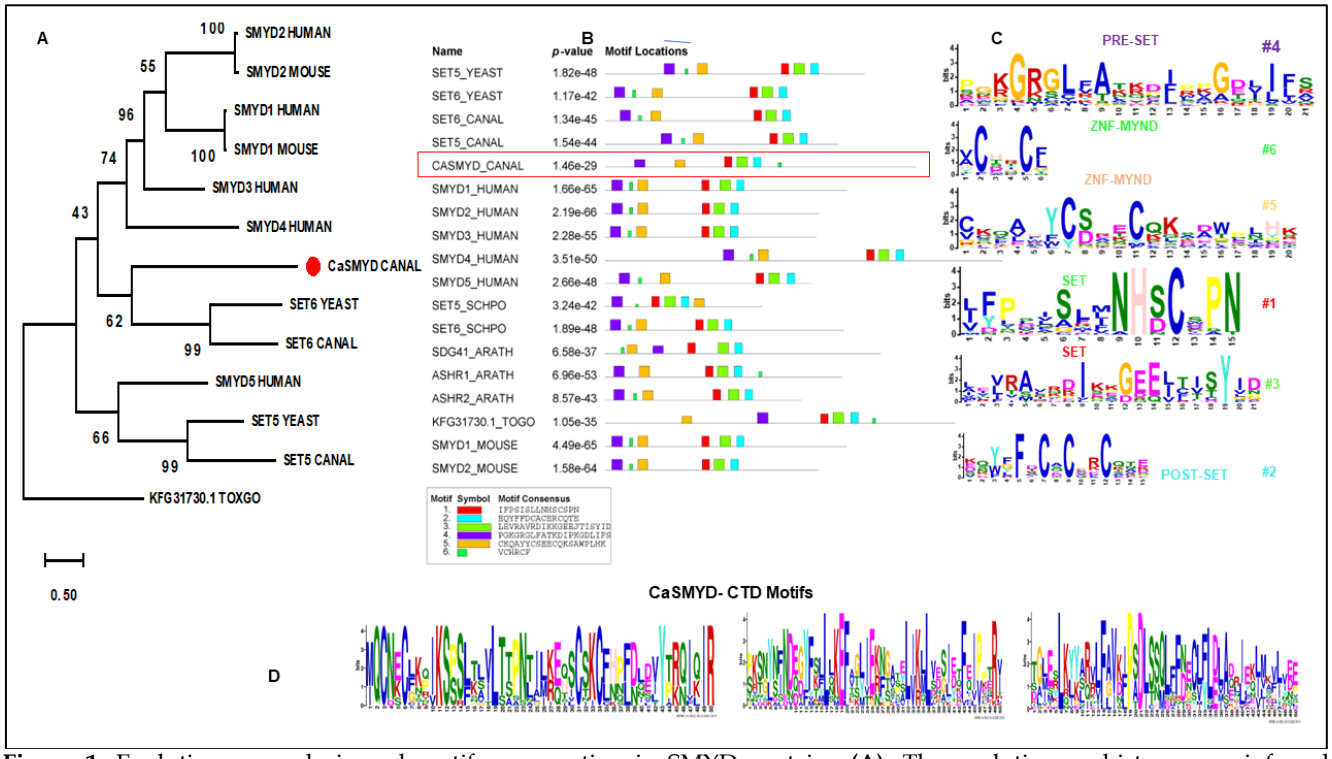

Figure 1: Evolutionary analysis and motif conservation in SMYD proteins **(A)**. The evolutionary history was inferred by using the Maximum Likelihood method and JTT matrix-based model. The accession numbers of the protein sequences are mentioned in Table 2. The percentage of trees in which the associated taxa cluster together is shown next to the branches. **(B)** MEME motif conservation across SMYND family members suggest the N terminal SET-MYND motif divergence. **(C)** Sequence logos for the motifs identified by MEME in panel B. **(D)** The C-terminal CX2CX24CX2C finger and other conserved motifs in *Candida* clade CaSMYD like proteins.

Various *Arabidopsis thaliana* and *Schizosaccharomyces pombe* SMYD KMTs were found in Uniprot blast **(Table 2),** however, no orthologs were detected in *Saccharomyces cerevisiae* or related yeasts in the Saccharomycetaceae family. High-identity sequences were identified among Saccharomycetales through blastp searches. The hits were identified in the genus *Candida, Lodderomyces, Spathaspora, Debaryomyces, Scheffersomyces,* and *Meyerozyma* belonging to the CUG-Ser1 clade. *Komagataella, Pichiaceae* and *Saccharomycetales incertae sedis* are the other clades having CaSMYD like proteins. Evolutionary analysis using Human, Mouse, yeast, and *Toxoplasma gondii* proteins identified CaSMYD to be an out group to the Set6 proteins of *Saccharomyces cerevisiae* and *Candia albicans*. CaSet5 & ScSet5 clustered with the human SMYD5 protein **(Figure 1(A))** while CaSMYD, ScSet6, and CaSet6 were out grouped into the human SYMD1-4 cluster **(Figure 1(A))**.

We extended sequence conservation studies with MEME analysis to indicate highly divergent SMYD domains **(Table 1).** SMYD domain architecture was found in several proteins, including CaSMYD, but with the lowest p-value **(Figure 1 (B))**, suggesting a divergent structure. CaSMYD domains included pre-SET (low identity), SET, and post-SET, domains, but CaSMYD lacked the zinc finger MYND domain C1X2C2X9C3X2C4X5C5X3C6X6H8X3C8 or its variant. The N-terminal ZNF-MYND motif (CX2C, motif of CaSMYD was after the post-SET (FXCXCX2C) (**motif#2 Figure.1(B))**. This was like the ASHR1 (*Arabidopsis thaliana*) and AKMT (*Toxoplasma gondii*) (**Figure.1 (C))** proteins**.** Indeed, the MEME motif analysis with yeast CaSMYD proteins identified the motif#6 to be C terminal part of a new C2C2 ZNF (CX2CX24CX2C) motif present in closely related *Candida* species (**Figure.1(D).** The position of this C2C2 ZNF in CaSMYD spanned amino acid 323 to 372**.** Also, two unique C-terminal domains (CTD motifs, **(Figure.1 (D))** were enriched in all non-saccharomyces yeast proteins. According to these results, CaSMYD protein contains SET-MYND domain, but its structure differs from Set5 and Set6 or other SMYD proteins. Therefore, *Candia albicans* CaSMYD protein contain uncharacterized SET-MYND domain, making it a non-canonical member of the SMYD family of SET MYND sub family of SET lysine methyltransferases. This report presents structural and functional annotations of diverged CaSMYD protein in comparison to other members of this family **(Table 3, and 4)**. This sub-family was present in several yeast clades but not in model organisms *S. cerevisiae* and *Schizosaccharomyces pombe* **(Figure. 2(A)).** The molecular weight was predicted to be approximately 73330.87 and reported to be a stable protein with an instability index of 39.90, like CaSet6, while CaSet5 is predicted to be unstable (instability index 50.67) (**Table 3**).

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 18(12): 1146-1153 (2022)

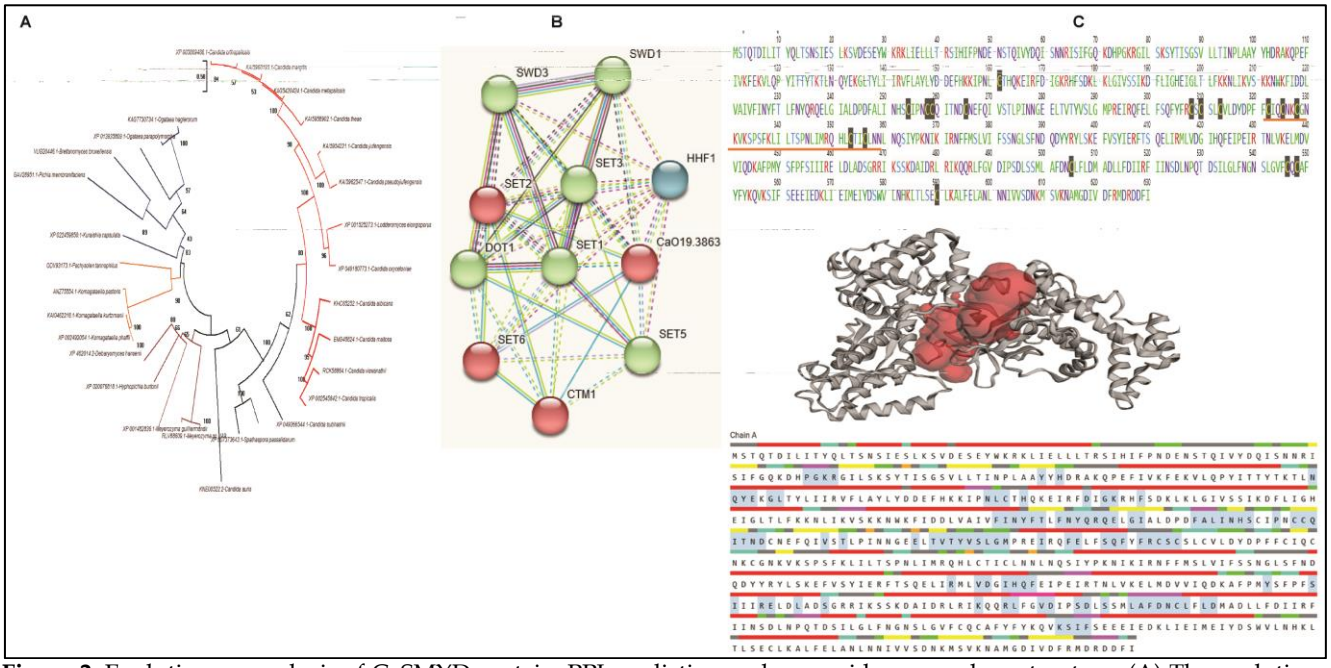

Figure 2: Evolutionary analysis of CaSMYD protein, PPI prediction and pre residue secondary structure. (A) The evolutionary history was inferred by using the Maximum Likelihood method JTT model on the MSA of yeast CaSMYD orthologs. The *Candida* clade is shown in red branches. The accession number and names are mentioned in the tree. (B) Prediction of the clustered PPI network of CaSMYD protein with nodes representing several methyltransferase and histone proteins. (C) Amino acids are colored based on their R group with Cysteine residues highlighted. The *Candida* clade specific C-terminal C2C2 finger sequence is underlined (Upper panel). The active site in the CaSMYD using the AlphaFold3D model is depicted with binding pocket residues shaded in gray and per residue secondary structure (Lower panel).

Table 3: The physiochemical, and structural features of *Candia albicans* SMYND proteins

Parameters	Outcomes		
Protein	A0A1D8PT54 (CaSmyd)	Q5A1M3 (CaSet5)	Q59VZ3 (CaSet6)
Amino acid count	630	473	379
Molecular weight	73330.87	53880.35	43918.02
Molecular formula	$C_{3334}H_{5202}N_{846}O_{952}S_{30}$	$C_{2374}H_{3765}N_{665}O_{722}S_{22}$	$C_{1991}H_{3036}N_{490}O_{594}S_{18}$
Theoretical pI	6.18	6.92	5.45
Instability index	39.90 (Stable)	50.67 (Unstable)	39.47 (Stable)
Aliphatic index	101.94	82.64	76.86
Grand average of hydropathicity	-0.053	-0.502	-0.403
InterPro	IPR046341: SET_dom_sf (24-352)); IPR001214: SET_dom (22-377)	IPR046341: SET_dom_sf (107-381) IPR001214: SET_dom (107-408)	IPR046341: SET_dom_sf (180-287), IPR001214: SET_dom (198-317)
SMART	set_7 - SM00317: set_7 (24-352)	SM000317 (107-387)	SM000317 (32-360)
CATH	SET-domain-G3DSA:2.170.270.10: SET domain (253-377)	SET-domain-G3DSA:2.170.270.10: SET domain (120-408), -Gene3D entry Model: 5v37A02 (156-318)	SET-domain-G3DSA:2.170.270.10: SET domain (198-317)
CDD	SET_SMYD - cd20071: SET_SMYD (279-372)	SET_SMYD-cd20071: SET_SMYD (335-407)	SET_SMYD - cd20071: SET_SMYD (221-313)
PROSITE	SET - PS50280: SET domain profile (24-346)	SET - PS50280: SET domain profile (107-381)	SET - PS50280: SET domain profile (180-287)
HHpred	3N71, E-value: 7.4e-29, Histone Lysine Methyltransferase Smyd1	4WUY, E-value: 1.9e-32, N-lysine methyltransferase SMYD2	6P7Z, E-value: 5.5e-32, SMYD3
PolyPhobius	Non-Cytoplasmic (1 - 533); Transmem. (534 -553) Cytoplasmic (554-630)	Non-Cytoplasmic.	Non-Cytoplasmic.
Yloc	Nucleus,64.03% small (0.18)	Nucleus 96.45%, very strong (1.00)	Nucleus 76.68% normal (0.65)
DeepLoc	Cytoplasm, Nucleus	Cytoplasm, Nucleus	Cytoplasm, Nucleus

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 18(12): 1146-1153 (2022)

	[0.7879;0.6321]; Nuclear export signal.	[0.6603;0.5850]; Nuclear localization signal, Nuclear export signal.	[0.6622;0.6818]; Nuclear export signal
MULoc Deep	Nucleus	Nucleus, Nucleolus, Chromosome	Nucleus, Chromosome
SignalP - 6.0	Non-secretory	Non-secretory	Non-secretory
CYSPRED (high 7 & above)	Bonding state 4 sites, Nonbonding state 9 Cysteine	Bonding state 4 sites, Nonbonding state 6 Cysteine	Bonding state 3 sites, Nonbonding state 7 Cysteine
DeepTMHMM	No Hits	No Hits	No Hits
HHrepID	322-340; 350-360	No repeats	No repeats
DeepCoils2	Low prob:	High prob.	No Coils
PREDICT PROTEIN: Meta-Disorder	1-2, 15-17, 19, 71	208,210-213, 251, 407-444; 451-462, 465-473	1-7, 134-140,142, 165, 167, 214-215, 219, 353, 371-379
PREDICT PROTEIN: Pro-NA Protein Binding	234-235; 244-45 252-244; 423-427;450, 501	26,27,28-29, 310-315, 317-328, 342-343, 382-384	382-384
PREDICT PROTEIN: Pro-NA DNA Binding	72-82; 255; 450-454; 609-613	62-70, 114-123, 154-161, 340-344, 366-370, 404-406	185-189, 210-215, 458-464
Disorder (Lambda Predict Protein)	1-4	1-25,417-473	1-9
Metal Binding (Lambda Predict Protein)	251,252,254, 310,313,325,328, 353 355	157,160,178,181,185,190,198,203,343-344,346,402,404,407	21,23,51,102-103,276,303-304,306,360,362,365
Small Molecule (Lambda Predict Protein)	76, 251,252,286	117-120,185,343-344, 380,398,400	21,23,51,102,103,276,303-304,337, 356,358
FoldSeek Top Hit (PDB 100) using 3Di/AA	7o2a_A (SMYD3); TM-Score: 0.49988; Identity 15.5%, Score 643, E=1.825e-10, Query Posi. 61-517 (630); Target Posi.1-424 (427)	4ynd_A (SMYD2), TM-Score: 0.67239; Identity 16.4%; Score 584, E= 3.932e-12; Query Posi. 103-412 (473); Target Posi 1-270 (428)	4wuy_A_A (SMYD2), TM-Score: 0.67239; Identity 16.2%; Score 617, E= 8.927e-13; Query Posi. 23-377 (379); Target Posi 2-268 (414)

A negative gravy value (-0.053) shows that the protein is more water soluble than CaSet5 and Caset6. The globular protein CaSMYD has a higher aliphatic index of 101.94, indicating thermostability over Caset5 and CaSet6. Interpro, SMART, CATH, and CDD databases identified SET domain profiles, but no other domain profiles matched in *Candida* proteins. Also, the HHpred hmm-hmm profile search yielded the highest CaSMYD hit with human HMT (Smyd1), whereas CaSet5 and CaSet6 profile-profile searches yielded the highest hits with human Smyd2 and Smyd3 **(Table 3)**. The *Candida albicans* SMYD KMT family comprises canonical CaSet5 & 6 and non-canonical CaSMYD proteins. CaSMYD proteins lack transmembrane domains, coils, or secretory signals, and are predicted to be found both in the nucleus and cytoplasm with a Nuclear Export Signal (NES). CaSMYD amino acids were predicted to bound macromolecules such as DNA, proteins, and small molecules, including metals (**Table 3**). Based on these findings, it is likely that the reported protein will have similar requirements and function as the other SMYD KMTs, which bind to SAM, zinc, and protein lysine as substrates. It is believed that protein intrinsically disordered regions are crucial structural and functional regulatory regions. CaSMYD was predicted to contain an N-terminal disordered region involved in protein binding. Among the three *Candida* SMYD methyltransferases, Caset5 was predicted to have a high confidence C-terminal large disorder region (400-473 aa). To increase the confidence of correct functional assignment to the uncharacterized proteins, the protein-protein interaction (PPI) network was evaluated for the CaSMYD protein annotation **(Figure 2 (B)).** Interacting partners of CaSMYD were SET2, SET6, and CTM1 in cluster 1 (PPI enrichment p-value: 6.03e-10). The predicted biological process was methylation (GO:0032259, FDR 0.00045); while the Molecular functions associated was histone-lysine n-methyltransferase (GO:0018024, FDR), Protein-lysine n-methyltransferase activity (GO:0016279, FDR 9.17e-05) and the methyltransferase activity (GO:0008168, FDR 9.17e-05). The predicted KEGG pathway was of lysine degradation (cal00310, FDR 6.98e-08) **(Figure 2(B))**. The secondary structures for amino acid residues were predicted in eight states. Of the CaSMYD protein's secondary structure, 43.02% are helices, 17.94% are extended helices, 33.97% are of random coils, and 5.08% are beta turns **(Figure 2 (C))**. For full-length proteins and domain sequences, AlphafFold machine learning model was used to generate the tertiary structure of SET and MYND proteins. Through CLUSTp 3.0, 108 amino acids were identified as being involved in the formation of the CaSMYD protein active site, based on the best ranked full-length models. The best active site had the Richard's solvent accessible surface area of 2121.012 $Å^2$ and solvent accessible volume of 1919.608 $Å^3$ **(Figure.2(C))**. We used the AlphFold models to identify homologs using reciprocal best structural hits and sequence similarities. To get the best structural alignment from PDB100 database, we used AlphFold models for three SMYD proteins in PDB files in FoldSeek server search. CaSMYD structure matched the SMYD3 protein with a low Tm score **(Table 4),** suggesting considerable divergence. We used the generated 3D models to determine the small molecule binding propensity of *Candida* SMYD KMTs. GalaxySite and AlphaFill webservers predicted the same ligand binding spectrum as the CaSMYD proteins for two canonical *Candida* SMYD proteins.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 18(12): 1146-1153 (2022)

Table 4: Predicted Ligand Binding to CaSMYD

Ligand Name	Drug Bank Name	Templates for protein-ligand complex	Binding Residues	Remarks
S-Adenosylmethionine (SAM)	Ademetionine	3QXY_B, 3RC0_A, 5CCL_A	53T 56V 67I 68F 69G 70Q 71K 77R 78G 159N	Co-factor
Sinefungin (SFG)	Sinefungin	3N71_A, 2H2J_C	56V 70Q 76K 77R 78G 156K 1248A 249L 250I 251N 252H 286Y 304F 306F, 59N 160L 161C,	Antifungal, Antiparasitic Agents
Lysine (LYS)	Lysine	1OZV_A	94I 98A 99A 100Y 101Y 156K 227N 228Y 229F	Approved nutraceutical and substrate of KMTs
(5-cyano-2'-{4-[2-(3-methyl-1H-indol-1-yl) ethyl] piperazin-1-yl}-N-[3-(pyrrolidin-1-yl) propyl] biphenyl-3-carboxamide (3UJ)	LLY-507	4WUY_A	247F 256P 260Q 284V 285T 287V 288S 289L 451S 452F 563E	Selective Inhibitor of Protein-lysine Methyltransferase SMYD2
N-cyclohexyl-N~3~-[2-(3,4-dichlorophenyl) ethyl]-N-(2-{[2-(5-hydroxy-3-oxo-3,4-dihydro-2H-1,4-benzoxazin-8-yl) ethyl] amino} ethyl)-beta-alaninamide (NH5)	AZ 505	3S7B_A	243L 244D 247F 248A 286Y 287V 288S 289L 290G 291M 295I 296R 299E 303Q	Oxazines/ Benzoxazines Class SMYD2 inhibitor.
5'-{[(3S)-3-amino-3-carboxypropyl] [3-(diethylamino) propyl] amino}-5'-deoxyadenosine (62X)	62X	5hi7.A	-	L-Peptide Linking. SMYD3 inhibitor.
S-Adenosyl-L-Homocysteine (SAH)	S-adenosyl-L-homocysteine	5hi7.A, 5CCL_A	-	Inhibitor, binder of Histamine N-methyltransferase

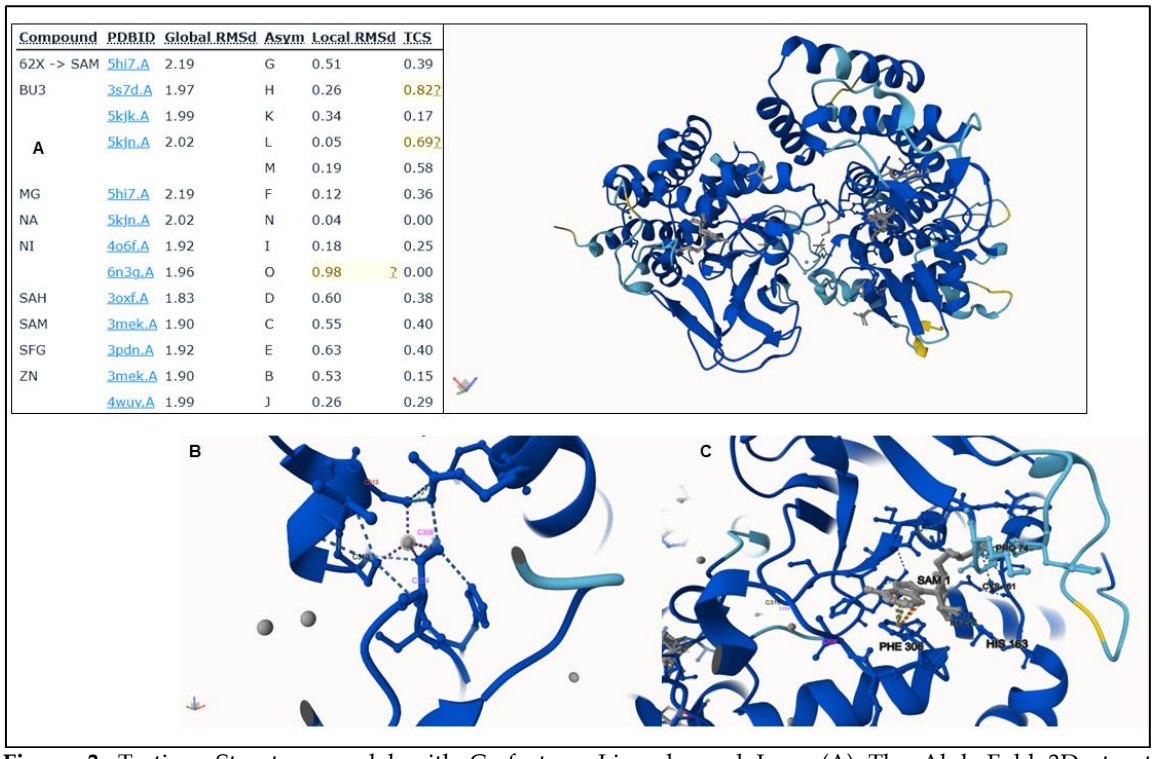

Figure 3: Tertiary Structure model with Co-factors, Ligands, and Ions. (A) The AlphaFold 3D structure of CaSYMD (UniProt ID: A0A1D8PT54) transplanted with different co-factors, ligands and ions based on AlphaFill algorithm. The compound names are mentioned in Table 4. The tertiary model is colored according to the pLDDT score, and the unfavorable transplant score in the table is highlighted in yellow. (B) Zinc ion coordination with Cysteine residues (C) S-adenosylmethionine (SAM) binding poses depicting amino acids and non-covalent interactions. H-bonds (blue dashed lines), Cation-Pi interaction (orange lines), and Pi stacking (Green lines).

Among different predicted CaSMYD ligands were, a cofactor S-adenosylmethionine (SAM), substrate lysine, and metal ions like zinc and nickel **(Figure 3 (B) & 3 (C))**. Furthermore, we also identified binding of multiple inhibitors of human SMYD2 and SMYD3 proteins bound to the CaSMYD active sites and structural elements. Molecules like sinefungin, 62X, NH5, and LLY-507, are predicted to bind CaSMYD (**Table 4**). Our manuscript describes the SET and MYND zinc finger KMTs of *Candida albicans* and proposes a new member of the Set5 and Set6 SMYD subfamily of methyltransferases. This SMYD member lacks a MYND zinc finger. Thus, we have identified a new lysine methyltransferase within the pathogen's genome utilizing well-established and recent tools for annotation of uncharacterized enzymes. We previously reported a *Komagataella phaffii* ortholog with the SMYD family of yeast **[37]**. The MYND-less subfamily of SMYD KMTs has been proposed based on analysis of phylogeny, conservation of the SET motif and MYND motif, binding residue prediction, and similarities between 3D structures and ligand binding data. As a result of this study, a solid foundation has been laid for future research into the biochemistry and molecular function of CaSMYD protein in *Candida albicans*.

Conclusions:

This study provides an understanding of *Candida albicans* SMYD methyltransferases, including CaSMYD (orf19.3863), a new member of the family. Protein sequence analysis, structural modelling, and structure-based docking studies were used to determine the potential role of this uncharacterized protein as a member of the protein lysine methyltransferase family.

Funding: This work was not supported by any financial assistance.

Conflict of Interest: None

References:

[1] Spellmon N *et al*. *Int J Mol Sci.* 2015 **16**:1406. [PMID: 25580534]
[2] Zhang Y *et al*. *Biomolecules* 2022 **12**:783. [PMID: 35740908]
[3] Jaiswal D *et al*. *Mol Cell Biol.* 2020 **40:**e00341. [PMID: 31685550]
[4] Rueda-Robles A *et al*. *Arch Biochem Biophys.* 2021 **712:**109040. [PMID: 34555372]
[5] Rubio-Tomás T. *Heliyon* 2021 **7:**e07387. [PMID: 34235289]
[6] Bernard BJ *et al*. *Clin Epigenetics* 2021 **13:**45. [PMID: 33637115]
[7] Bottino C *et al*. *Cancers (Basel)* 2020 **12:**142. [PMID: 31935919]
[8] Singh RP *et al*. *J Biol Chem*. 2011 **286:**25154. [PMID: 21592964]
[9] Potter SC *et al*. *Nucleic Acids Res*. 2018 **46:**W200. [PMID: 29905871]
[10] Altschul S F *et al*. *J Mol Biol*. 1990 **215**:403. [PMID:2231712]
[11] Jones P *et al*. *Bioinformatics* 2014 **30:**1236. [PMID: 24451626]
[12] Söding J *et al*. *Nucleic Acids Res*. 2005 **33:**W244. [PMID: 15980461]
[13] Gasteiger E *et al*. *Nucleic Acids Res.* 2003 **31:**3784. [PMID: 12824418]
[14] Käll L *et al*. *Nucleic Acids Res*. 2007 **35:**W429. [PMID: 17483518]
[15] Briesemeister S *et al*. *Nucleic Acids Res*. 2010 **38:**W497. [PMID: 20507917]
[16] Thumuluri V *et al*. *Nucleic Acids Res*. 2022 **50:**W228. [PMID: 35489069]
[17] Jiang Y *et al*. *Comput Struct Biotechnol J*. 2021 **19:**4825. [PMID: 34522290]
[18] Teufel F *et al*. *Nat Biotechnol*. 2022 **40:**1023. [PMID: 34980915]
[19] https://dtu.biolib.com/DeepTMHMM
[20] Buchan DWA *et al*. *Nucleic Acids Res*. 2019 **47:**W402. [PMID: 31251384]
[21] Gabler F *et al*. *Curr Protoc Bioinformatics* 2020 **72:**e108. [PMID: 33315308]
[22] Edgar R C *et al*. *Nucleic Acids Res*. 2004 **32:**1792. [PMID:15034147]
[23] Zhang D. *et al*. *Mol Ecol Resour*. 2020 **20:**348. [PMID: 31599058]
[24] Tamura K *et al*. *Mol Biol Evol*. 2021 **38:**3022. [PMID: 33892491]
[25] Bailey TL *et al*. *Nucleic Acids Res*. 2009 **37:**W202. [PMID: 19458158]
[26] Bernhofer M *et al*. *Nucleic Acids Res*. 2021 **49:**W535. [PMID: 33999203]
[27] Fariselli P *et al. Proteins* 1999 **36:**340. [PMID: 10409827]
[28] Szklarczyk D *et al*. *Nucleic Acids Res*. **49:**10800. [PMID: 33237311]
[29] https://embed.predictprotein.org/
[30] Jumper J *et al*. *Nature* 2021 **596:**583. [PMID: 34265844]
[31] Mirdita M *et al*. *Nat Methods*. 2022 **19:**679. [PMID: 35637307]
[32] Hekkelman ML *et al*. *Nat Methods* 2022 **10.1038**/s41592-022-01685-y. [PMID: 36424442]
[33] https://search.foldseek.com/search
[34] Sehnal D. *et al. Nucleic Acids Res.* 2021 **49:**W43. [PMID: 33956157]
[35] Ko J *et al*. *Nucleic Acids Res*. 2012 **40:**W294. [PMID: 22649060]
[36] Tian W *et al*. *Nucleic Acids Res*. 2018 **46:**W363. [PMID: 29860391]
[37] Gogoi RK *et al*. *Bioinformation* 2019 **15:**542. [PMID: 31719763]