Clues to reaction specificity in PLP-dependent fold type I enzymes of monosaccharide biosynthesis

Jaya Srivastava and Petety V. Balaji

Department of Biosciences and Bioengineering, Indian Institute of Technology Bombay, Powai, Mumbai 400076, India

Correspondence: Jaya Srivastava, jaya_srivastava@iitb.ac.in

Keywords: motifs, structure-function, bioinformatics, sugar aminotransferases

List of abbreviations

NSAT: Nucleotide Sugar Aminotransferase
NSD: Nucleotide Sugar Dehydratase
HMM: Hidden Markov Model
PWM: Position Weight Matrix
MSA: Multiple Sequence Alignment
SDR: Short-chain Dehydrogenase Reductase
ROC: Receiver Operator Characteristic

Includes Supplementary_files containing Tables S1-S4, Figures S1-S7 and Supplementary data.xlsx

Abstract

Novel functions can emerge in an enzyme family while conserving catalytic mechanism, motif or fold. PLP-dependent enzymes have evolved into seven fold types and catalyse diverse reactions using the same mechanism for the formation of external aldimine. Nucleotide sugar aminotransferases (NSATs) and dehydratases (NSDs) belong to fold type I and mediate the biosynthesis of several monosaccharides. NSATs use diverse substrates but are highly selective to the C3 or C4 carbon to which amine group is transferred. Factors responsible for reaction specificity in NSDs are known but remain unexplored in NSATs. Profile HMMs were able to identify NSATs but could not capture reaction specificity. A search for discriminating features led to the discovery of a sequence motif that is located near the pyranose binding site suggesting their role in imparting reaction specificity. Using a position weight matrix for this motif, we were able to assign reaction specificity to a large number of NSATs. Residues which upon mutation could convert NSD to NSAT have been reported in literature and we deduced that these are not conserved. This suggested the occurrence of non-generic family specific mutations underlying the evolution of dehydratases. Inferences from this analysis set way for future experiments that can shed light on mechanisms of functional diversification in enzymes of fold type I.

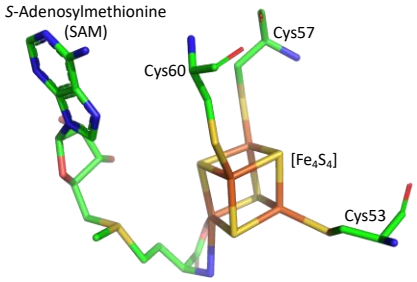
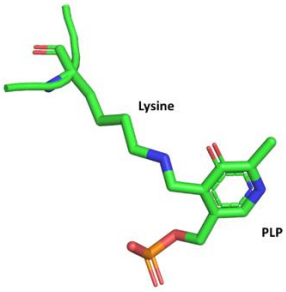
1. INTRODUCTION

Monosaccharides are the building blocks of carbohydrates and constitute the 'third alphabet of life'. This alphabet has many more building blocks than even the protein alphabet [1] and this is a consequence of glycodiversification [2], [3]. Structural heterogeneity of monosaccharides is a functional requirement, especially among prokaryotes as they adapt to varying environmental conditions by evolving novel monosaccharides and glycans [2]. Glycan building blocks differ from each other primarily in the configuration and functional group substituents of carbon atoms [4], [5]. These variations are brought about by only a handful of enzyme families as revealed by an analysis of the biosynthesis pathways of 55 monosaccharides [6]. These enzyme families are involved in other biological processes as well. For instance, a sugar 3-dehydrogenase that catalyses the biosynthesis of UDP-Man2NAc3NAcA [7] belongs to the GFO/IDH/MocA family and this family contains enzymes of glycolysis and galactose metabolism also [8]. Clearly, organisms recruit existing enzymes, tweak substrate specificity and reaction mechanism through sequence and structural modifications and thereby generate novel enzymes.

The conserved core of an enzyme family can be defined by 1) a few residues, such as cofactor binding lysine in PLP dependent enzymes [9], 2) a motif, as that in Radical SAM superfamily [10], or 3) a conserved fold, as that in Short chain dehydrogenase Reductases (SDRs) [11] (Table 1). PLP-dependent enzymes use the same mechanism for the formation of external aldimine even though they catalyze different types of reactions [12]. They have been grouped into seven fold types based on a comparison of their 3D structures (Figure 1) [13] thereby reflecting catalytic strategies that have independently evolved among PLP dependent enzymes [12]. Nucleotide sugar aminotransferases (NSATs) and nucleotide sugar dehydratases (NSDs) belong to fold type I [14].

This group also contains enzymes that catalyze several other types of reactions [15] (Figure 1). NSATs and NSDs take part in the biosynthesis of at least 19 monosaccharides (Figure S1) and secondary metabolites [16] [17].

Table 1: Features conserved at various levels among functionally diverse enzymes and illustrative examples

Conserved feature	Representative 3D rendering
FOLD: SDR superfamily Representative example from [18] Rendering: PDB ID 2GN4. Region highlighted in pink shows the core Rossmann fold domain of SDR superfamily **Associated functions**: Dehydratase, dehydrogenase, enolase, epimerase, isomerase, reductase, etc.	
SEQUENCE MOTIF: Cys-X_3-Cys-X_2-Cys Representative example from [19] Rendering: PDB id 1R30 **Associated functions**: Dehydrogenase, lyase, methyltransferase, etc.	
RESIDUE: Lysine in vitamin B_6 dependent enzymes Representative example from [20] Rendering: PDB ID 3NYU **Associated functions**: Decarboxylase, dehydrase, ligase, lyase, mutase, phosphorylase, transaminase, etc.	

NSATs convert keto-sugars to deoxy amino sugars and differ from each other in three ways: (i) substrate specificity, (ii) whether amino group is transferred to the C3 or C4 carbon of the pyranose ring, and (iii) the configuration of the chiral center formed as a consequence of amino group transfer. Some of the experimentally characterized NSATs (Figure 2, Table S1) are specific to the

nature and orientation of substituents on pyranose ring [21] whereas others exhibit broad substrate specificity [22], [23]. However, they are specific towards the position and orientation of amine group installation [21], [23], [24]. The conformation of the backbone is highly conserved among NSATs with known 3D structures (Figure 3A) despite of sequence divergence (Figure S2).

NSDs utilize similar substrates as NSATs but catalyse dehydration in place of transamination. ColD (a NSD) utilizes the same substrate as Per (a C4 NSAT) but catalyzes dehydration as opposed to transamination [25]. These catalytic differences among NSATs and NSDs have emerged by keeping the overall structure intact (Figure 3B). Even the extent of sequence similarity among functional variants of NSATs and NSDs is so comparable (Figure S2) that it is not possible to distinguish between these two families solely based on sequence similarity. Clearly, reaction specificity is determined by as yet unknown "minor" differences in primary structures. In this study, we have identified features that discriminate C3 NSATs from C4 NSATs using profile Hidden Markov Models (HMMs) and position weight matrices (PWM). In addition, we have analysed the extent of conservation of residues proposed in earlier studies to be responsible for dehydratase activity among NSDs. Through these analyses, we provide clues to sequence features imparting reaction specificity in C3 NSATs, C4 NSATs and NSDs.

2. METHODS

Databases, software and online servers: Protein sequences were obtained from UniProtKB (Release 2021_1) [26] and 3D structures from the Protein Data Bank [27]. Sequences which had one or more of the characters B, O, J, U, X and Z were ignored. The CATH database (v4.2) was used for identifying proteins

which share the same fold [28]. Pairwise sequence comparisons were performed using BLASTp online server [29]. Pairwise identity matrix was obtained using MUSCLE webserver [30]. Multiple sequence alignments (MSAs) were done using the MAFFT command line application [31]. 3D structures were visualized, analyzed and compared using PyMol [32]. HMMER3 [33] was used to build profile HMMs. Electron density maps of ligand structures were visualized using WinCOOT [34].

Datasets and profile HMMs: Experimentally characterized NSATs and NSDs were obtained from literature; reviewed homologs were taken from SwissProt (Table S1). Datasets used and HMMs generated in this study are described in Tables S4. Bit score thresholds for $NSAT_{HMM}$ and NSD_{HMM} were set based on Receiver Operator Characteristic (ROC) curves generated using hits from TrEMBL under the assumption that annotations provided in TrEMBL are correct. The procedure used to generate ROC curves is described in detail elsewhere [6], [35]. Briefly, ROC curves were generated by calculating the number of true positives, false positives and false negatives for various values of bit score threshold (Figure S3). TrEMBL entries which did not have molecular function annotation were ignored. Possible thresholds chosen from the inflection point of ROC curve were validated by scanning the sequences of fold type I PLP-dependent enzyme FoldTypeI_exp dataset, Table S4). Based on this procedure, thresholds for $NSAT_{HMM}$ and NSD_{HMM} were set to be 310 and 300 bits, respectively.

ROC curves could not be used to set thresholds for $C3_NSAT_{HMM}$ and $C4_NSAT_{HMM}$ since annotations in TrEMBL just mention NSAT without specifying if the protein is a C3 NSAT or C4 NSAT. To overcome this hurdle, an alternative approach was used whereby the TrEMBL database was searched

with these two profiles separately by using the default value suggested by the HMMER software as threshold (viz., E-value = 10). A bit score scatter plot of hits common to both profiles was obtained and on the basis of this plot, thresholds for both C3_NSAT$_{HMM}$ and C4_NSAT$_{HMM}$ were set to 400 bits (Figure 4A).

Identification of sequence motifs: The command line application of STREME [36] was used to identify sequence motifs enriched in a sequence family with reference to a set of "control" sequences. The dataset C4_hits (Table S4) was used as control sequence dataset to discover motifs in C3_hits and vice versa. Position-specific amino acid frequency matrices were obtained from STREME and converted into a PWM of log odds score viz., $\log_2 \frac{f_{obs}}{f_{exp}}$, where f_{obs} and f_{exp} are observed and expected amino acid frequencies in a given position of the MSA. Expected frequencies were fetched from TrEMBL (Release 2021_1). For a given sequence, the region that has the highest log odds score was assumed to be the best possible match for the motif; in fact, the highest scoring motifs were verified to be the motifs of interest using in-house python scripts. Steps leading to motif discovery using profile HMMs are schematically represented in Figure 5A.

Phylogenetic analysis: Phylogenetic analysis was performed using MEGA7; sequences were first aligned by MUSCLE using default values for all the parameters and then creating a phylogenetic tree by Maximum Likelihood method [37]. Sequences from dataset FoldTypeI_exp with >=80% identity were removed, and the resulting 208 enzymes, including PglE from *Camplyobacter jejuni* (PDB ID: 4ZTC), a C4 NSAT that is structurally identical to other C4 NSATs

that is yet to be updated in CATH, were used for phylogenetic analysis. The tree was visualized using iTOL [38].

3. RESULTS

Identifying homologs of NSATs using profile HMMs: Scanning TrEMBL using C3_NSAT$_{HMM}$ and C4_NSAT$_{HMM}$ resulted in ~14000 hits for each profile. Twenty-one hits satisfied the threshold of both the profiles (viz., 400 bits) (Supplementary data, worksheet: common_hits). Among sequences scoring below 400 against both profiles, 32,344 were hits to NSAT$_{HMM}$. These sequences are predicted to be NSATs (because they are hits of NSAT$_{HMM}$) but we could not resolve if these are C3 NSATs or C4 NSATs, or neither of these two. Similarly, we could not resolve if 21 common hits are C3 NSATs or C4 NSATs, or have both activities. Thus, our inability to set a bit score threshold that is sensitive enough to identify C3 NSATs and C4 NSATs without compromising specificity meant that an alternative approach is required to discriminate C3 NSATs and C4 NSATs from each other.

Identifying motifs that discriminate C3 NSATs and C4 NSATs from each other: We used a MSA of sequences used to build C3_NSAT$_{HMM}$ and C4_NSAT$_{HMM}$ to identify differences in sequence features of these two groups of enzymes. The MSA showed the presence of a 7-residue motif C-terminus to the active site lysine in C3 NSATs (Figure 6A). The role of either the first glycine residue conserved in C3 NSATs but not in C4 NSATs or of any other residues of this 7-residue motif has not been investigated so far. Is the sequence information in this motif adequate to discriminate C3 NSATs from C4 NSATs? To check this, C3_hits and C4_hits (Table S4) were used to find motifs enriched in C3 NSATs

in comparison to C4 NSATs and vice versa using STREME[1]. The most conserved motif among C3_hits is a 15-mer that encompasses the aforementioned 7-residue motif and found in 99.9% of input sequences[2] (Figure 6). The most conserved motif among C4 NSATs is a 12-mer located at the same site as the conserved motif in C3_hits (Figure 6). This motif is found in all but one C4_exp sequences. Visual inspection of the sequence logos confirmed the distinctiveness of these motifs (Figure 6). Both motifs are present in the active site of NSATs (Figure 6D). We postulate that these motifs influence the positioning of the substrate in the binding site for C3 or C4 amine installation. These motifs are henceforth referred to as C3_motif and C4_motif, respectively.

Setting bit score threshold for PWMs: Two PWMs viz., $C3_{PWM}$ and $C4_{PWM}$, were generated using the MSAs of C3_motif and C4_motif, respectively. The entire length of each and every sequence in C3_hits and C4_hits was scanned by both these PWMs. The highest scoring 15- and 12-mer in each sequence was deemed to be the desired motif. The score of the highest scoring 15-/12-mer of a sequence was assigned as the score for that sequence. The distribution of scores showed the near-absence of 15-/12-mers that have positive scores when C3_hits are scanned using $C4_{PWM}$ and when C4_hits are scanned using $C3_{PWM}$ (Figure 7).

[1] we used hits of profile HMMs instead of only the experimentally characterized sequences for this purpose with the expectation that a larger number of input sequences may lead to higher sensitivity of motif search

[2] whether a 12-mer that excluded positions 1-3 of the 15-mer motif can be a potential C3_motif was also probed. It was found that the highest scoring 12-mers in six sequences from C3_hits were not from the expected position, but adjacent to it. Hence, this 12-mer motif was not considered for any further analysis

When C3_hits were scanned using $C3_{PWM}$, four sequences scored <0 bits; three of these contain deletions in the C3_motif, including deletion of active site lysine; the fourth appears to be a C4 NSAT based on the high score (35 bits) when scanned against C4_motif. When C4_hits are scanned using $C3_{PWM}$, five sequences score >0 bits; four among these score >30 bits when scanned against $C4_{PWM}$ and the identity of the fifth sequence could not be ascertained based on $C3_{PWM}$ or $C4_{PWM}$.

When C4_hits are scanned using $C4_{PWM}$, all the sequences score >6 bits except for one which scores 0.07 bits, because of a two-residue insertion. However, when C3_hits are scanned using $C4_{PWM}$, 32 sequences score > 0 bits. One among these appears to be a C4 NSAT as it scores <0 bits against $C3_{PWM}$. Three others score in the range 6-15 bits against $C4_{PWM}$ but > 35 bits against $C3_{PWM}$; their identities could not be resolved using motif search. The rest score in the range 0-3.5 bits against $C4_{PWM}$ but > 30 bits against $C3_{PWM}$. Based on these data, thresholds for $C3_{PWM}$ and $C4_{PWM}$ were set to 3 and 4 bits, respectively. Hence, with four exceptions, motif search captures differences among C3 NSATs and C4 NSATs that could not be identified using profile HMMs.

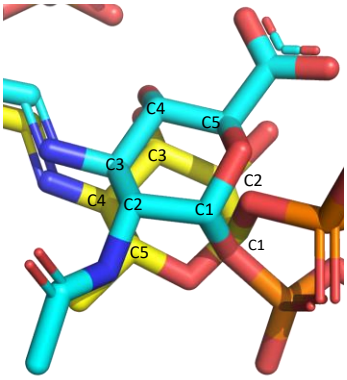
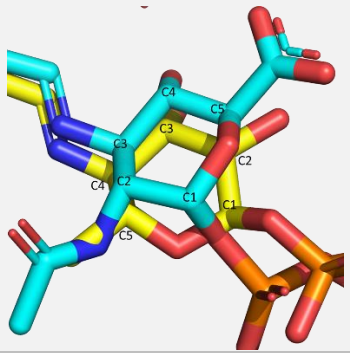
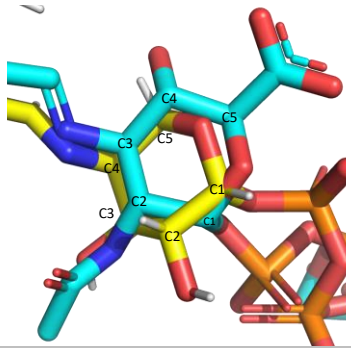
<u>Resolving common hits</u>: As mentioned earlier, 21 sequences are hits to both $C3_NSAT_{HMM}$ and $C4_NSAT_{HMM}$ (Supplementary data, worksheet: common_hits). They were scanned using $C3_{PWM}$ and $C4_{PWM}$. 19 sequences are distinctly identified as C3 NSAT or C4 NSAT based on PWM scores. However, two proteins with accession numbers A0A2G9NIF3 and A0A1W9P6M8 satisfy the log-odds threshold of both motifs. Both profile HMM and PWM-based searches are not effective in resolving the specificities of these two proteins. Is it possible that these two enzymes have broad reaction specificity?

Sorting out C3 NSATs and C4 NSATs among hits of $NSAT_{HMM}$: Data from the scanning of C3_hits and C4_hits using the two PWMs show that C3_motif and C4_motif are unique to respective families of proteins. Hence, a two-step screening wherein the amino acid sequence of a protein is (i) first scanned using $C3_NSAT_{HMM}$ and $C4_NSAT_{HMM}$, and (ii) then by $C3_{PWM}$ and $C4_{PWM}$ is proposed for the functional annotation of C3 NSATs and C4 NSATs (Figure 5B). The 32,344 sequences that score below the thresholds of $C3_NSAT_{HMM}$ and $C4_NSAT_{HMM}$ but are hits to $NSAT_{HMM}$ were analyzed for the presence of C3_motif and C4_motif. 25,135 were predicted as C4 NSATs and 5,887 as C3 NSATs. It can be seen that C3 NSATs score less than C4 NSATs against $C4_NSAT_{HMM}$; however, the converse is not true i.e., C4 NSATs do not necessarily score less than C3 NSATs against $C3_NSAT_{HMM}$ (Figure 4B). 1073 and 249 sequences, respectively, score less than and higher than the thresholds of both PWMs. We are unable to identify their function based on the two-step process flow proposed herein. One possible reason is that the process is not sensitive enough. Another possibility is that some of these sequences have broad reaction specificity. Alternatively, these sequences may have lost NSAT activity and acquired novel function(s) due to sequence changes that are too subtle to be detected by HMMs and PWMs.

Additional factors governing product identity - clues from 3D structures: We superposed the 3D structures of three C3 NSATs and three C4 NSATs bound to substrates in the form of external aldimine (Table 2). The backbone conformation is highly similar in these six proteins (Figure 3A). In addition, the binding modes and conformations of ligands are also similar as can be seen from the overlap in the position of PLPs and of the nucleotides in the six complexes (Figure S5A). However, PLP forms aldimine with C3 keto or C4 keto

group and this difference is accommodated by conformation change around the pyrophosphate group (Figure S5A). The conformation of ligands is identical in C3 NSATs (Figure S5B); in the other three, i.e., C4 NSATs, the pyranose ring either flips or shifts (Figure S5C, S5D). A tyrosine residue H-bonds with the N-acetyl / -OH group at C2 in all three crystal structures and this residue is conserved in C3_hits (Table 2). No such H-bond is observed in the three crystal structures of C4 NSATs and this may be the reason for the observed shift/flip of the pyranose ring. Multiple sequence alignment of C4_hits shows that the homologous position is a tyrosine in only ~40% of the sequences; in rest of the sequences, it is a tryptophan or phenylalanine. It has been proposed that ring flipping leads to variations in stereochemical outcome at C4 [39]. If that indeed is the case, it may explain the occurrence of both axial and equatorial configuration among C4 NSATs but only one among C3 NSATs. Experimental verification is required to confirm the role of these surrounding residues as well as the discriminating sequence motifs that were used for PWM-based motif search, in determining the position of amine installation.

Table 2: NSATs with bound substrates in the form of external aldimine used for structural analysis[¶]

Protein[‡]	PDB id	Pyranose[@]	Position aminated	Conformation Nucleotide sugar (relative to each other)[§]	Key interaction[†]
WbpE	3NYT	$^4C_1$	C3	Used as the reference conformation for the purpose of comparison	Tyr309...NAc@C2
WlarG	5U21	$^4C_1$	C3	Same as that in WbpE	Tyr306...NAc@C2
QdtB	3FRK	$^4C_1$	C3	Same as that in WbpE	Tyr310...NAc@C2
DesI	2PO3	$^4C_1$	C4	**FLIPPED** 	Tyr309 (of WbpE) is replaced by Phe330 and hence, no H-bond.
PerA	3DR4	$^4S_2$	C4	**FLIPPED** 	Tyr309 (of WbpE) is replaced by Phe318 and hence, no H-bond.
ArnB	4OCA	$^4C_1$	C4	**SHIFTED** 	Tyr309 (of WbpE) is replaced by Phe330 and hence, no H-bond.

[¶] 3D structures with bound substrates in external aldimine form are also available for WecE (PDB id 4ZAH), and PseC (PDB id 2FNU), but these were not considered for analysis – the ligand conformation was deemed uncertain based on a visual inspection of the fit of the structure to the electron density (Figure S6, Prof. P. Bhaumik, personal communication).

[‡] Other details [name of the organism, UniProt accession number, length, PubMed Id of the article which describes its experimental characterization and gene name(s)] of these enzymes are given in Supplementary data, Worksheet: Exp and reviewed sequences.

[@] Conformation of the pyranose ring in the bound substrate.

[§] In case of DesI and ArnB, only the pyranose rings are shown; complete structures are shown in Figure S5.
[†] Tyr309 of WbpE is conserved in 99.6% of C3_hits. The homologous residue is a tyrosine in only 40% of C4_hits.

Features that distinguish NSATs and NSDs from each other: NSDs are involved in the biosynthesis of NDP-3,6-dideoxy sugars (Table S1). The overall reaction involves C3 deoxygenation, and is achieved by a single enzyme in the pathways for the biosynthesis of GDP-L-colitose and TDP-forosamine [25], [40] but by two enzymes in the pathway for the biosynthesis of CDP-linked 3,6-dideoxy sugars [41]. NSDs are unique among PLP-dependent enzymes in having a histidine in place of the lysine that forms Schiff's base with PLP [42]. The need for lysine is done away with in NSDs, since they do not form internal aldimine but instead release ammonia in the reaction [25]. Histidine has been proposed to play a role in general acid-base catalysis [41], and its replacement with lysine stalls dehydratase reaction [43]. The reaction mechanism leading to the formation of the ketimine intermediate is the same in both NSDs and NSATs [43] and features unique to NSDs and to NSATs drive 3-deoxygenation or transamination of this intermediate.

In addition to this Lys/His replacement, other mutations are also proposed to contribute to difference in the reaction catalysed (Table 3). We sought to analyse the extent of conservation in these five sites, i.e., positions corresponding to Asp194 (Site 1), Tyr217 (Site 2) His220 (site 4), and Phe345 (Site 5) (*Yersinia pseudotuberculosis* E1 dehydratase numbering) and Ser187 (Site 3) (*Escherichia coli* ColD numbering) in C3_hits, C4_hits and NSD_hits that were obtained from TrEMBL. Tabulation of analysis is provided in Table 3. Active site lysine and histidine (Site 4) are conserved in NSATs and NSDs respectively. At Site 1, glutamate is found in both NSDs and C4 NSATs, while

aspartate is only found in NSDs. Similarly, tyrosine is found in all three groups. Occupancy at site 3 suggests that serine is not exclusive to NSDs and this is also evident from experiments which find that the S187N single mutant retains dehydratase activity [43]. Site 5 is occupied by residues other than histidine in both C3 NSATs and C4 NSATs. All of these observations suggest that multiple sites contribute to the conversion of transaminase to dehydratase activity, and that altered reaction specificity cannot be generalised to specific sites. We also identified an asparagine residue near the N-terminus of NSDs (Site 6) that lies in the PLP binding site (Asn60 in WbdK, PDB ID: 3B8X, figure not shown), and is conserved in ~95% of NSDs. Its proximity to PLP suggests its possible involvement in modulating chemical properties of the ketimine intermediate; however experimental validation will shed more light on its properties.

Table 3: Analysis of the extent of conservation of residues implicated in distinguishing transamination from dehydration

Site[¥]	Site 1[¶]		Site 2[¶]		Site 3[‡]		Site 4[¶‡@]		Site 5[¶]		Site 6	
Enzyme family	NSAT	NSD	NSAT	NSD	NSAT	NSD	NSAT	NSD	NSAT	NSD	NSAT	NSD
Residue	His	Asp[§]	His[£]	Tyr	Asn[†]	Ala	Lys	His	His[£]	Phe	Met	Asn
Position[$]	163	194	185	217	185	219	188	220	297	345	67	90
C3_hits * (12811 sequences)	Gln: 12712 His: 41 Asp: 0		Tyr: 7703 Phe: 4954 His: 80		Ser: 4529 Gly: 4112 Thr: 3397 Ala: 603 Asn: 4		Lys: 12799 His: 0		His: 9605 Ala: 1389 Asn: 1108 Phe: 0		Leu: 11247 Iso: 1273 Val: 223 Asn: 25	
C4_hits * (12101 sequences)	His: 6868 Glu: 3602 Gln: 1482 Asp: 0		His: 6005 Asn: 2614 Tyr: 2462		Asn: 4126 Ile: 2363 Val: 2782 Thr: 2049 Ser: 71 Ala: 23		Lys: 12100 His: 0		His: 7469 Trp: 3414 Phe: 219		Leu: 8770 Met: 1886 Iso: 1354 Asn: 0	
NSD_hits * (4018 sequences)	Asp: 3173 Glu: 844 His: 0		Tyr: 3249 Phe: 764 His: 5		Ala: 2303 Ser: 856 Pro: 822 Asn: 0		His: 4018 Lys: 0		Phe: 3783 His: 27		Asn: 3849 Asp: 86 Leu: 8	

[¥] Sites 1 to 5 were chosen based on site directed mutagenesis studies reported in literature. Site 6 is chosen based on visual analysis of MSA (this study).

[$]Positions in the primary structure to which the six sites correspond to and the residue at these positions correspond to *Salmonella Typhimurium* ArnB (StArnB) for NSAT and *Yersinia pseudotuberculosis* E1 dehydratase for NSD.

* The number of sequences in which residue conservation analysis is done is given in parenthesis. For each site and enzyme family, the number of sequences in which the most frequently occurring residues / the residue in the reference proteins (mentioned in the column heading) occur are shown.

[¶] Data for Sites 1, 2, 4 and 5 are for *Yersinia pseudotuberculosis* E1 dehydratase, an NSD. Wild type residues were replaced by residues at the homologous position in *Salmonella Typhimurium* ArnB, a C4 NSAT. The four site (D194H-Y217H-H220K-F345H) mutant showed aminotransferase activity [41].

[‡] Data for Sites 3 and 4 are for *Escherichia coli* GDP-4-keto 6-deoxy mannose 3-dehydratase (ColD), an NSD. Wild type residues were replaced by residues at the homologous position in GDP-perosamine synthase (PerA), a C4 NSAT. The two site (S187A-H188K) mutant showed aminotransferase activity [43]. The residue at position 187 is Ala in *Yersinia pseudotuberculosis* E1 dehydratase.

[@] Lysine in this site forms Schiff's base with PLP. This lysine is replaced by histidine in NSDs.

[§] Asp194 has been proposed to stabilize the ketimine intermediate through a salt bridge with 3-oxygen of the cofactor [44]. This residue is replaced by histidine in StArnB.

[£] NSATs have histidine at Sites 2 and 5; these are not conserved in NSDs. The role of these two histidines in transamination is not known.

[†] Asn185 in PerA H-bonds with the pyranose ring oxygen [45] and this H-bond is absent in ColD.

4. Discussion

Fold type I PLP-dependent enzymes constitute a family of enzymes wherein sequence and structural variations have led to functional diversification within a conserved fold (Figure 1). An earlier 3D structure-based phylogenetic study of fold type I PLP-dependent enzymes revealed a 'sugar AT family' consisting of NSATs, NSDs, non-nucleotide sugar aminotransferases and AHBA synthase [46]. We obtained a similar sugar AT family of 22 enzymes from a sequenced based phylogenetic analysis (Figure S7). Functions associated with other clusters also are largely conserved within a clade. The similarities in the 3D structure-based and sequence-based trees implies that sequence changes that lead to functional changes are reflected in 3D structures despite of fold conservation. What features govern reaction selectivity in NSATs and NSDs?

Aside from sequence and structural similarity among NSATs, conserved PLP and nucleotide group bound conformations suggest that differences are confined to regions that interact with pyranose ring. Using a combination of profile HMMs and motif based PWMs, we located a sequence motif in the pyranose binding site that differs among C3 NSATs and C4 NSATs. Structural comparisons also led to the discovery of a conserved tyrosine-aided H-bond among C3 NSATs that interacts and stabilizes the pyranose group. Tyrosine or similar residues are absent in ~60% of C4 NSAT homologs, suggesting that absence of stabilizing interactions may lead to flipping or shifting of pyranose ring to orient C4 keto group for amination.

Other enzymes in the 'sugar AT family' contain NSDs that utilize nucleotide linked sugars but catalyse dehydration instead. A hallmark of NSDs is the replacement of active site lysine with histidine, though its replacement with lysine does not convert it to an aminotransferase, but merely stalls its native

activity. Additional residues have been shown to alter a dehydratase to an aminotransferase, but they are not strictly conserved among NSDs. 'Sugar AT family' also contains aminotransferases that utilize sugars not linked to nucleotides. NtdA, a 3-oxo-glucose-6-phosphate aminotransferase [47], and KdnA, a 8-amino-3,8-dideoxy-alpha-D-*manno*-octulosonate aminotransferase [48] are examples of non-nucleotide sugar aminotransferases. These enzymes harbour variations in active site that accommodate nucleotide group. For example, the supposed nucleotide binding pocket of NtdA is occupied by side chains of Leu70 and Tyr246 [47]. In RbmB, a 2-deoxy-*scyllo*-inosose aminotransferase [49], a longer loop protrudes Phe336 into the nucleotide binding pocket, which along with Gln190 obstructs the binding of nucleotide group. Collectively, these observations suggest the presence of non-generic family specific features which cannot be deduced using sequence-based approaches.

The folding space of polypeptides is finite [50]–[54]. Consequently, the same fold can accommodate multiple functions using diverse mechanisms such as point mutations, indels, recruitment of additional domains, oligomerization, etc. [55]. Several factors such as gene loss, enzyme promiscuity, biological networks, etc., are proposed to drive neo- and subfunctionalization [56]–[58]. NSATs and NSDs are examples of functional families wherein a few mutations lead to alterations in reaction specificity. In the absence of enough experimentally characterized proteins that can capture these mutations, automated function prediction algorithms that provide functional annotation using sequence similarity are likely to lead to incomplete or erroneous annotations at the level of substrate and reaction specificity. Such limitations are reflected in the inability of C3_NSAT$_{HMM}$ and C4_NSAT$_{HMM}$ to identify C3 NSATs and C4 NSATs from hits of NSAT$_{HMM}$. Limitations of profile HMMs are

overcome by combining family specific motif search approach. Thus, in an attempt to provide clues to reaction specificity among NSATs and NSDs, we have underscored the necessity of manual curation in protein function prediction and suggested approaches to overcome limitations of automated annotation algorithms.

PLP-dependent enzymes host a wide range of catalytic reactions and are highly relevant biocatalysts [59]. Minor alterations in the active site leading to functional diversity are being routinely engineered for biotechnological applications [60], [61]. With experimental validation, features highlighted in this study that contribute to reaction specificities in NSATs and its homologs may be useful in engineering efficient enzymes for the biosynthesis of natural products.

Author contributions

PVB conceptualized research, JS performed analysis, PVB and JS cowrote the paper.

Acknowledgements

The authors are thankful to Professor Prasenjit Bhaumik for providing inputs on structural analysis. Jaya Srivastava is thankful to the Council of Scientific and Industrial Research, Government of India, for a research fellowship [number 09/087/(0877)/2017-EMR-I].

5. References

[1] H.-J. Gabius and J. Roth, “An introduction to the sugar code,” *Histochem. Cell Biol.*, vol. 147, no. 2, pp. 111–117, Feb. 2017, doi: 10.1007/s00418-016-1521-9.

[2] R. J. Mostowy and K. E. Holt, “Diversity-Generating Machines: Genetics of Bacterial Sugar-Coating,” *Trends Microbiol.*, vol. 26, no. 12, pp. 1008–1021, 2018, doi: 10.1016/j.tim.2018.06.006.

[3] C. J. Thibodeaux, C. E. Melançon, and H. Liu, “Natural-product sugar biosynthesis and enzymatic glycodiversification,” *Angew. Chem. Int. Ed. Engl.*, vol. 47, no. 51, pp. 9814–9859, 2008, doi: 10.1002/anie.200801204.

[4] H.-J. Gabius, “The sugar code: Why glycans are so important,” *Biosystems*, vol. 164, pp. 102–111, Feb. 2018, doi: 10.1016/j.biosystems.2017.07.003.

[5] P. H. Seeberger, “Monosaccharide Diversity,” in *Essentials of Glycobiology*, 3rd ed., A. Varki, R. D. Cummings, J. D. Esko, P. Stanley, G. W. Hart, M. Aebi, A. G. Darvill, T. Kinoshita, N. H. Packer, J. H. Prestegard, R. L. Schnaar, and P. H. Seeberger, Eds. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press, 2015. Accessed: Dec. 08, 2020. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK453086/

[6] J. Srivastava, P. Sunthar, and P. V. Balaji, “The glycan alphabet is not universal: a hypothesis,” *Microb Genom*, vol. 6, no. 11, Nov. 2020, doi: 10.1099/mgen.0.000452.

[7] J. B. Thoden and H. M. Holden, “Structural and functional studies of WlbA: A dehydrogenase involved in the biosynthesis of 2,3-diacetamido-2,3-dideoxy-D-mannuronic acid,” *Biochemistry*, vol. 49, no. 36, pp. 7939–7948, Sep. 2010, doi: 10.1021/bi101103s.

[8] H. Taberman, T. Parkkinen, and J. Rouvinen, “Structural and functional features of the NAD(P) dependent Gfo/Idh/MocA protein family oxidoreductases,” *Protein Sci.*, vol. 25, no. 4, pp. 778–786, Apr. 2016, doi: 10.1002/pro.2877.

[9] E. F. Oliveira, N. M. F. S. A. Cerqueira, P. A. Fernandes, and M. J. Ramos, “Mechanism of formation of the internal aldimine in pyridoxal 5’-phosphate-dependent enzymes,” *J Am Chem Soc*, vol. 133, no. 39, pp. 15496–15505, Oct. 2011, doi: 10.1021/ja204229m.

[10] J. L. Vey and C. L. Drennan, “Structural insights into radical generation by the radical SAM superfamily,” *Chem Rev*, vol. 111, no. 4, pp. 2487–2506, Apr. 2011, doi: 10.1021/cr9002616.

[11] M. Da Costa *et al.*, “Structure-function relationships in NDP-sugar active SDR enzymes: Fingerprints for functional annotation and enzyme engineering,” *Biotechnol Adv*, vol. 48, p. 107705, Feb. 2021, doi: 10.1016/j.biotechadv.2021.107705.

[12] G. Schneider, H. Käck, and Y. Lindqvist, “The manifold of vitamin B6 dependent enzymes,” *Structure*, vol. 8, no. 1, pp. R1-6, Jan. 2000, doi: 10.1016/s0969-2126(00)00085-x.

[13] “B6db home : Main Menu.” http://bioinformatics.unipr.it/cgi-bin/bioinformatics/B6db/home.pl (accessed Jul. 13, 2021).

[14] K. Fesko, D. Suplatov, and V. Švedas, “Bioinformatic analysis of the fold type I PLP-dependent enzymes reveals determinants of reaction specificity in l-threonine aldolase from Aeromonas jandaei,” *FEBS Open Bio*, vol. 8, no. 6, pp. 1013–1028, Jun. 2018, doi: 10.1002/2211-5463.12441.

[15] J. Catazaro, A. Caprez, A. Guru, D. Swanson, and R. Powers, “Functional evolution of PLP-dependent enzymes based on active-site structural similarities,” *Proteins*, vol. 82, no. 10, pp. 2597–2608, Oct. 2014, doi: 10.1002/prot.24624.

[16] A. J. Romo and H. Liu, “Mechanisms and structures of vitamin B(6)-dependent enzymes involved in deoxy sugar biosynthesis,” *Biochim. Biophys. Acta*, vol. 1814, no. 11, pp. 1534–1547, Nov. 2011, doi: 10.1016/j.bbapap.2011.02.003.

[17] S. C. Timmons and J. S. Thorson, “Increasing carbohydrate diversity via amine oxidation: aminosugar, hydroxyaminosugar, nitrososugar, and nitrosugar biosynthesis in bacteria,” *Curr Opin Chem Biol*, vol. 12, no. 3, pp. 297–305, Jun. 2008, doi: 10.1016/j.cbpa.2008.03.017.

[18] N. Ishiyama *et al.*, “Structural studies of FlaA1 from Helicobacter pylori reveal the mechanism for inverting 4,6-dehydratase activity,” *J Biol Chem*, vol. 281, no. 34, pp. 24489–24495, Aug. 2006, doi: 10.1074/jbc.M602393200.
[19] G. L. Holliday *et al.*, “Atlas of the Radical SAM Superfamily: Divergent Evolution of Function Using a ‘Plug and Play’ Domain,” *Methods Enzymol*, vol. 606, pp. 1–71, 2018, doi: 10.1016/bs.mie.2018.06.004.
[20] G. T. Dow, M. Gilbert, J. B. Thoden, and H. M. Holden, “Structural investigation on WlaRG from Campylobacter jejuni: A sugar aminotransferase,” *Protein Sci*, vol. 26, no. 3, pp. 586–599, Mar. 2017, doi: 10.1002/pro.3109.
[21] A. Larkin and B. Imperiali, “Biosynthesis of UDP-GlcNAc(3NAc)A by WbpB, WbpE, and WbpD: enzymes in the Wbp pathway responsible for O-antigen assembly in Pseudomonas aeruginosa PAO1,” *Biochemistry*, vol. 48, no. 23, pp. 5446–5455, Jun. 2009, doi: 10.1021/bi900186u.
[22] M. Useglio, S. Peirú, E. Rodríguez, G. R. Labadie, J. R. Carney, and H. Gramajo, “TDP-L-megosamine biosynthesis pathway elucidation and megalomicin a production in Escherichia coli,” *Appl. Environ. Microbiol.*, vol. 76, no. 12, pp. 3869–3877, Jun. 2010, doi: 10.1128/AEM.03083-09.
[23] S. Hwang, Z. Li, Y. Bar-Peled, A. Aronov, J. Ericson, and M. Bar-Peled, “The biosynthesis of UDP-d-FucNAc-4N-(2)-oxoglutarate (UDP-Yelosamine) in Bacillus cereus ATCC 14579: Pat and Pyl, an aminotransferase and an ATP-dependent Grasp protein that ligates 2-oxoglutarate to UDP-4-amino-sugars,” *J. Biol. Chem.*, vol. 289, no. 51, pp. 35620–35632, Dec. 2014, doi: 10.1074/jbc.M114.614917.
[24] J. B. Thoden, C. Schäffer, P. Messner, and H. M. Holden, “Structural analysis of QdtB, an aminotransferase required for the biosynthesis of dTDP-3-acetamido-3,6-dideoxy-alpha-D-glucose,” *Biochemistry*, vol. 48, no. 7, pp. 1553–1561, Feb. 2009, doi: 10.1021/bi8022015.
[25] J. Alam, N. Beyer, and H. Liu, “Biosynthesis of colitose: expression, purification, and mechanistic characterization of GDP-4-keto-6-deoxy-D-mannose-3-dehydrase (ColD) and GDP-L-colitose synthase (ColC),” *Biochemistry*, vol. 43, no. 51, pp. 16450–16460, Dec. 2004, doi: 10.1021/bi0483763.
[26] UniProt Consortium, “UniProt: a worldwide hub of protein knowledge,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D506–D515, Jan. 2019, doi: 10.1093/nar/gky1049.
[27] H. M. Berman *et al.*, “The Protein Data Bank,” *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235–242, Jan. 2000, doi: 10.1093/nar/28.1.235.
[28] I. Sillitoe *et al.*, “CATH: increased structural coverage of functional space,” *Nucleic Acids Res*, vol. 49, no. D1, pp. D266–D273, Jan. 2021, doi: 10.1093/nar/gkaa1079.
[29] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, Oct. 1990, doi: 10.1016/S0022-2836(05)80360-2.
[30] R. C. Edgar, “MUSCLE: multiple sequence alignment with high accuracy and high throughput,” *Nucleic Acids Res*, vol. 32, no. 5, pp. 1792–1797, 2004, doi: 10.1093/nar/gkh340.
[31] K. Katoh and D. M. Standley, “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability,” *Molecular Biology and Evolution*, vol. 30, no. 4, pp. 772–780, Apr. 2013, doi: 10.1093/molbev/mst010.
[32] “PyMOL | pymol.org.” https://pymol.org/2/ (accessed May 18, 2021).
[33] S. R. Eddy, “A new generation of homology search tools based on probabilistic inference,” *Genome Inform*, vol. 23, no. 1, pp. 205–211, Oct. 2009.
[34] P. Emsley and K. Cowtan, “Coot: model-building tools for molecular graphics,” *Acta Crystallogr D Biol Crystallogr*, vol. 60, no. Pt 12 Pt 1, pp. 2126–2132, Dec. 2004, doi: 10.1107/S0907444904019158.
[35] J. Srivastava, R. Hembrom, A. Kumawat, and P. V. Balaji, “Searching sequence databases for functional homologs using profile HMMs: how to set bit score thresholds?,” *bioRxiv*, p. 2021.06.24.449764, Jun. 2021, doi: 10.1101/2021.06.24.449764.

[36] T. L. Bailey, “STREME: Accurate and versatile sequence motif discovery,” *Bioinformatics*, Mar. 2021, doi: 10.1093/bioinformatics/btab203.
[37] S. Kumar, G. Stecher, and K. Tamura, “MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets,” *Mol Biol Evol*, vol. 33, no. 7, pp. 1870–1874, Jul. 2016, doi: 10.1093/molbev/msw054.
[38] I. Letunic and P. Bork, “Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation,” *Nucleic Acids Research*, vol. 49, no. W1, pp. W293–W296, Jul. 2021, doi: 10.1093/nar/gkab301.
[39] E. S. Burgie and H. M. Holden, “Molecular architecture of DesI: a key enzyme in the biosynthesis of desosamine,” *Biochemistry*, vol. 46, no. 31, pp. 8999–9006, Aug. 2007, doi: 10.1021/bi700751d.
[40] L. Hong, Z. Zhao, C. E. Melançon, H. Zhang, and H. Liu, “In vitro characterization of the enzymes involved in TDP-D-forosamine biosynthesis in the spinosyn pathway of Saccharopolyspora spinosa,” *J. Am. Chem. Soc.*, vol. 130, no. 14, pp. 4954–4967, Apr. 2008, doi: 10.1021/ja0771383.
[41] P. Smith, P.-H. Szu, C. Bui, H. Liu, and S.-C. Tsai, “Structure and Mutagenic Conversion of E1 Dehydrase: at the Crossroads of Dehydration, Aminotransfer and Racemization,” *Biochemistry*, vol. 47, no. 24, pp. 6329–6341, Jun. 2008, doi: 10.1021/bi702449p.
[42] P. D. Cook and H. M. Holden, “GDP-4-keto-6-deoxy-D-mannose 3-dehydratase, accommodating a sugar substrate in the active site,” *J. Biol. Chem.*, vol. 283, no. 7, pp. 4295–4303, Feb. 2008, doi: 10.1074/jbc.M708893200.
[43] P. D. Cook, R. L. Kubiak, D. P. Toomey, and H. M. Holden, “Two site-directed mutations are required for the conversion of a sugar dehydratase into an aminotransferase,” *Biochemistry*, vol. 48, no. 23, pp. 5246–5253, Jun. 2009, doi: 10.1021/bi9005545.
[44] B. W. Noland *et al.*, “Structural studies of Salmonella typhimurium ArnB (PmrH) aminotransferase: a 4-amino-4-deoxy-L-arabinose lipopolysaccharide-modifying enzyme,” *Structure*, vol. 10, no. 11, pp. 1569–1580, Nov. 2002, doi: 10.1016/s0969-2126(02)00879-1.
[45] P. D. Cook, A. E. Carney, and H. M. Holden, “Accommodation of GDP-linked sugars in the active site of GDP-perosamine synthase,” *Biochemistry*, vol. 47, no. 40, pp. 10685–10693, Oct. 2008, doi: 10.1021/bi801309q.
[46] Y. Itoh, M. J. Bröcker, S. Sekine, D. Söll, and S. Yokoyama, “Dimer-dimer interaction of the bacterial selenocysteine synthase SelA promotes functional active-site formation and catalytic specificity,” *J Mol Biol*, vol. 426, no. 8, pp. 1723–1735, Apr. 2014, doi: 10.1016/j.jmb.2014.01.003.
[47] K. E. van Straaten, J. B. Ko, R. Jagdhane, S. Anjum, D. R. J. Palmer, and D. A. R. Sanders, “The structure of NtdA, a sugar aminotransferase involved in the kanosamine biosynthetic pathway in Bacillus subtilis, reveals a new subclass of aminotransferases,” *J Biol Chem*, vol. 288, no. 47, pp. 34121–34130, Nov. 2013, doi: 10.1074/jbc.M113.500637.
[48] T. R. Zachman-Brockmeyer, J. B. Thoden, and H. M. Holden, “Structures of KdnB and KdnA from Shewanella oneidensis: Key Enzymes in the Formation of 8-Amino-3,8-Dideoxy-d-Manno-Octulosonic Acid,” *Biochemistry*, vol. 55, no. 32, pp. 4485–4494, Aug. 2016, doi: 10.1021/acs.biochem.6b00439.
[49] T. R. Zachman-Brockmeyer, J. B. Thoden, and H. M. Holden, “The structure of RbmB from Streptomyces ribosidificus, an aminotransferase involved in the biosynthesis of ribostamycin,” *Protein Sci*, vol. 26, no. 9, pp. 1886–1892, Sep. 2017, doi: 10.1002/pro.3221.
[50] R. D. Schaeffer and V. Daggett, “Protein folds and protein folding,” *Protein Eng Des Sel*, vol. 24, no. 1–2, pp. 11–19, 2011, doi: 10.1093/protein/gzq096.
[51] X. Liu, B. Lv, and W. Guo, “The size distribution of protein families within different types of folds,” *Biochem Biophys Res Commun*, vol. 406, no. 2, pp. 218–222, Mar. 2011, doi: 10.1016/j.bbrc.2011.02.020.

[52] S. Govindarajan, R. Recabarren, and R. A. Goldstein, “Estimating the total number of protein folds,” *Proteins*, vol. 35, no. 4, pp. 408–414, Jun. 1999.
[53] R. Kolodny, L. Pereyaslavets, A. O. Samson, and M. Levitt, “On the universe of protein folds,” *Annu Rev Biophys*, vol. 42, pp. 559–582, 2013, doi: 10.1146/annurev-biophys-083012-130432.
[54] G. Yang, C. M. Miton, and N. Tokuriki, “A mechanistic view of enzyme evolution,” *Protein Sci*, vol. 29, no. 8, pp. 1724–1747, Aug. 2020, doi: 10.1002/pro.3901.
[55] S. D. Copley, “Evolution of new enzymes by gene duplication and divergence,” *FEBS J.*, vol. 287, no. 7, pp. 1262–1283, Apr. 2020, doi: 10.1111/febs.15299.
[56] A. L. Juárez-Vázquez *et al.*, “Evolution of substrate specificity in a retained enzyme driven by gene loss,” *Elife*, vol. 6, p. e22679, Mar. 2017, doi: 10.7554/eLife.22679.
[57] M. E. Glasner, D. P. Truong, and B. C. Morse, “How enzyme promiscuity and horizontal gene transfer contribute to metabolic innovation,” *FEBS J*, vol. 287, no. 7, pp. 1323–1342, Apr. 2020, doi: 10.1111/febs.15185.
[58] C. Roth *et al.*, “Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms,” *J Exp Zool B Mol Dev Evol*, vol. 308, no. 1, pp. 58–73, Jan. 2007, doi: 10.1002/jez.b.21124.
[59] J. F. Rocha, A. F. Pina, S. F. Sousa, and N. M. F. S. A. Cerqueira, “PLP-dependent enzymes as important biocatalysts for the pharmaceutical, chemical and food industries: a structural and mechanistic perspective,” *Catalysis Science & Technology*, vol. 9, no. 18, pp. 4864–4876, 2019, doi: 10.1039/C9CY01210A.
[60] F. J. Fernandez *et al.*, “Structure and mechanism of a cysteine sulfinate desulfinase engineered on the aspartate aminotransferase scaffold,” *Biochim Biophys Acta*, vol. 1824, no. 2, pp. 339–349, Feb. 2012, doi: 10.1016/j.bbapap.2011.10.016.
[61] F. Steffen-Munsberg *et al.*, “Bioinformatic analysis of a PLP-dependent enzyme superfamily suitable for biocatalytic applications,” *Biotechnol Adv*, vol. 33, no. 5, pp. 566–604, Oct. 2015, doi: 10.1016/j.biotechadv.2014.12.012.

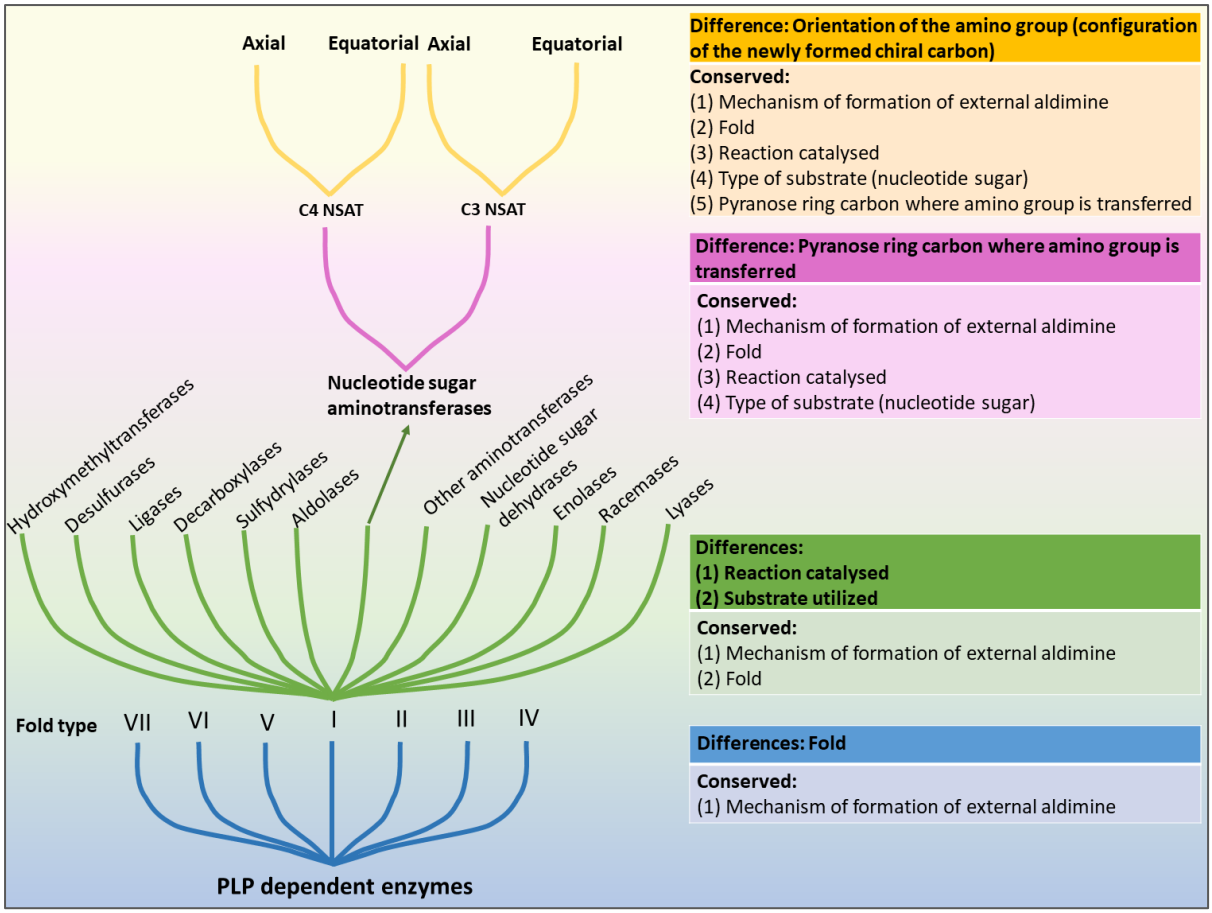
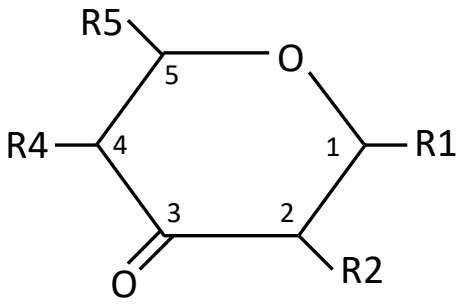


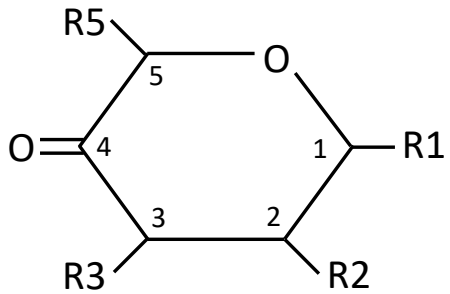
Figure 1: Neofunctionalization in PLP dependent enzymes leading to the evolution of specialized functions. Nucleotide sugar aminotransferases and dehydratases adopt fold type I. 'Other aminotransferases' include non-nucleotide linked aminotransferases as well as those utilizing distinct substrates such as amino-acids. Based on the pyranose ring carbon which is aminated, NSATs are further classified as C3 NSATs and C4 NSATs. These, in turn, can install amine either axially or equatorially (assuming $^{4}C_{1}$ conformation) (Table S2 and S3).

(A)



Gene name	Pyranose conformation	R1	R2	R4	R5	C3-NH_2
QdtB, WlarG, TylB	$^4C_1$	TDP/dTDP (Ax)	OH (Eq)	OH (Eq)	CH_3 (Eq)	(Eq)
FdtB, WlarG, RavAMT	$^4C_1$	TDP/dTDP (Ax)	OH (Eq)	OH (Ax)	CH_3 (Eq)	(Eq)
DesV	$^4C_1$	TDP/dTDP (Ax)	OH (Eq)	H	CH_3 (Eq)	(Eq)
EvaB, MegDII	$^4C_1$	TDP/dTDP (Ax)	H	=O	CH_3 (Eq)	(Eq)
WbpE	$^4C_1$	UDP (Ax)	NAc (Eq)	OH (Eq)	COOH (Eq)	(Eq)

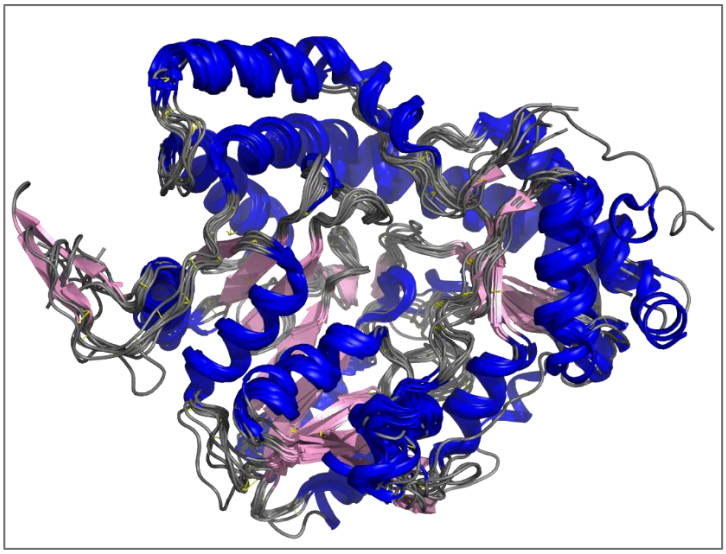
(B)



Gene name	Pyranose conformation	R1	R2	R3	R5	C4-NH_2
DesI, VioA	$^4C_1$	TDP/dTDP (Ax)	OH (Eq)	OH (Eq)	CH_3 (Eq)	(Eq)
WecE	$^4C_1$	TDP/dTDP (Ax)	OH (Eq)	OH (Eq)	CH_3 (Eq)	(Ax)
LegC	$^4C_1$	GDP (Ax)	NAc (Eq)	OH (Eq)	CH_3 (Eq)	(Eq)
WeeJ, PglC, PglE, EpsN	$^4C_1$	UDP (Ax)	NAc (Eq)	OH (Eq)	CH_3 (Eq)	(Eq)
Pat	$^4C_1$	UDP (Ax)	NAc (Eq)	OH (Eq)	CH_3 (Eq)	(Ax)
PseC, Pam	$^1C_4$	UDP (Eq)	NAc (Ax)	OH (Ax)	CH_3 (Eq)	(Eq)
Per, PerA, RfbE	$^4C_1$	GDP (Ax)	OH (Ax)	OH (Eq)	CH_3 (Eq)	(Eq)
ArnB	$^4C_1$	UDP (Ax)	OH (Eq)	OH (Eq)	H	(Ax)
SpnR	$^4C_1$	TDP/dTDP (Ax)	H	H	CH_3 (Eq)	(Eq)

Figure 2: Composite figures showing the reactants and products of C3 NSATs (A) and C4 NSATs (B). Eq and Ax denote equatorial and axial, respectively. The orientation of the -NH_2 group in the product is shown in the right-most column. The conformation of the pyranose ring has been so chosen that the substituent at C5 (i.e., the highest numbered chiral carbon in the pyranose ring) is oriented equatorially. Schematic structures of reactants and products are given in Tables S2 and S3.

(A)



(B)

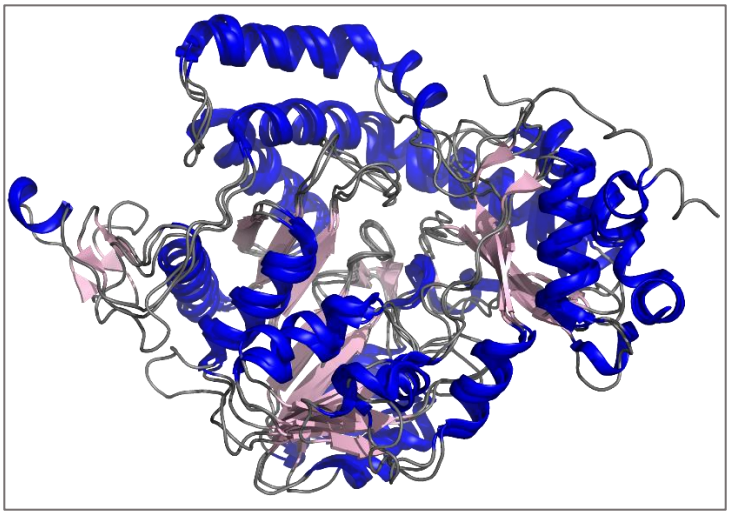


Figure 3A: Structural superimposition of four C3 NSATs and six C4 NSATs. PDB ids, substrates and orientation of the amino group in the product are given below.

C3 NSATs		
3NYU	UDP-2-acetamido-2-deoxy-3-oxo-glucuronate	Equatorial ($^4C_1$)
5U21	(d)TDP-3-keto-6-deoxy-galactose/glucose	Equatorial ($^4C_1$)
3FRK	(d)TDP-3-keto-6-deoxy-glucose	Equatorial ($^4C_1$)
2OGA	(d)TDP-3-keto-3,4,6-trideoxy-D-*erythro*-hexopyranose	Equatorial ($^4C_1$)
C4 NSATs		
4ZAH	(d)TDP-4-keto-6-deoxy-glucose	Axial ($^4C_1$)
2PO3	(d)TDP-4-keto-6-deoxy-glucose	Equatorial ($^4C_1$)
3DR4	GDP-4-keto-6-deoxy-mannose	Equatorial ($^4C_1$)
2FNU	UDP-2-acetamido-2,6-dideoxy-beta-L-*arabino*-4-hexulose	Equatorial ($^4C_1$)
1MDO	UDP-beta-L-*threo*-pentapyranos-4-ulose	Equatorial ($^4C_1$)
4ZTC	UDP-4-keto-6-deoxy-GlcNAc	Equatorial ($^4C_1$)

Figure 3B: Structural superimposition of a C3 NSAT (3NYT), a C4 NSAT (2PO3) and a NSD (2GMS)
Other details (UniProt accession number, length, organism, gene / protein name and PubMed id) are given in Supplementary data, Worksheet: Exp and reviewed sequences.

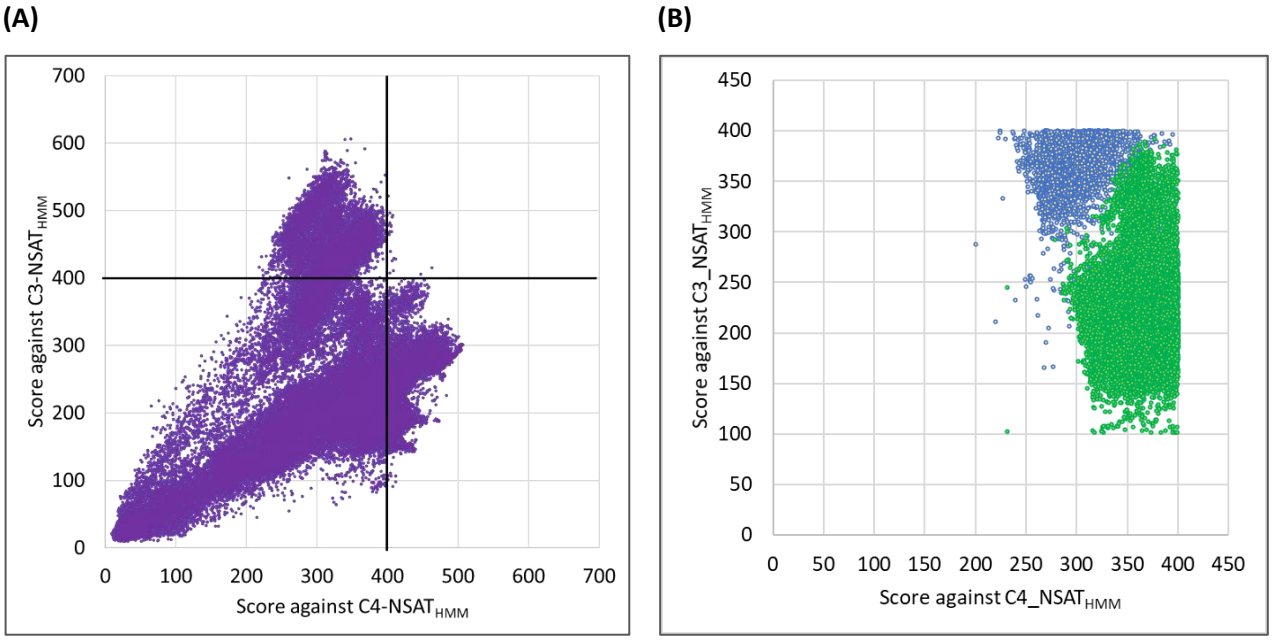


Figure 4: (A) Distribution of bit scores of hits common to both C3_NSAT$_{HMM}$ and C4_NSAT$_{HMM}$ from TrEMBL using default threshold suggested by the HMMER software. As can be seen, some proteins are "high-scoring" for one profile while being "low-scoring" for the other profile. This formed the basis on which the threshold was set 400 bits for both the profiles (B) Comparison of bit score and PWM log-odds score prediction. Green dots represent sequences that are predicted as C4 NSATs by motif search, while blue dots are predicted as C3. Note: this plot does not include 1,322 sequences for which PWM-based prediction was inconclusive.

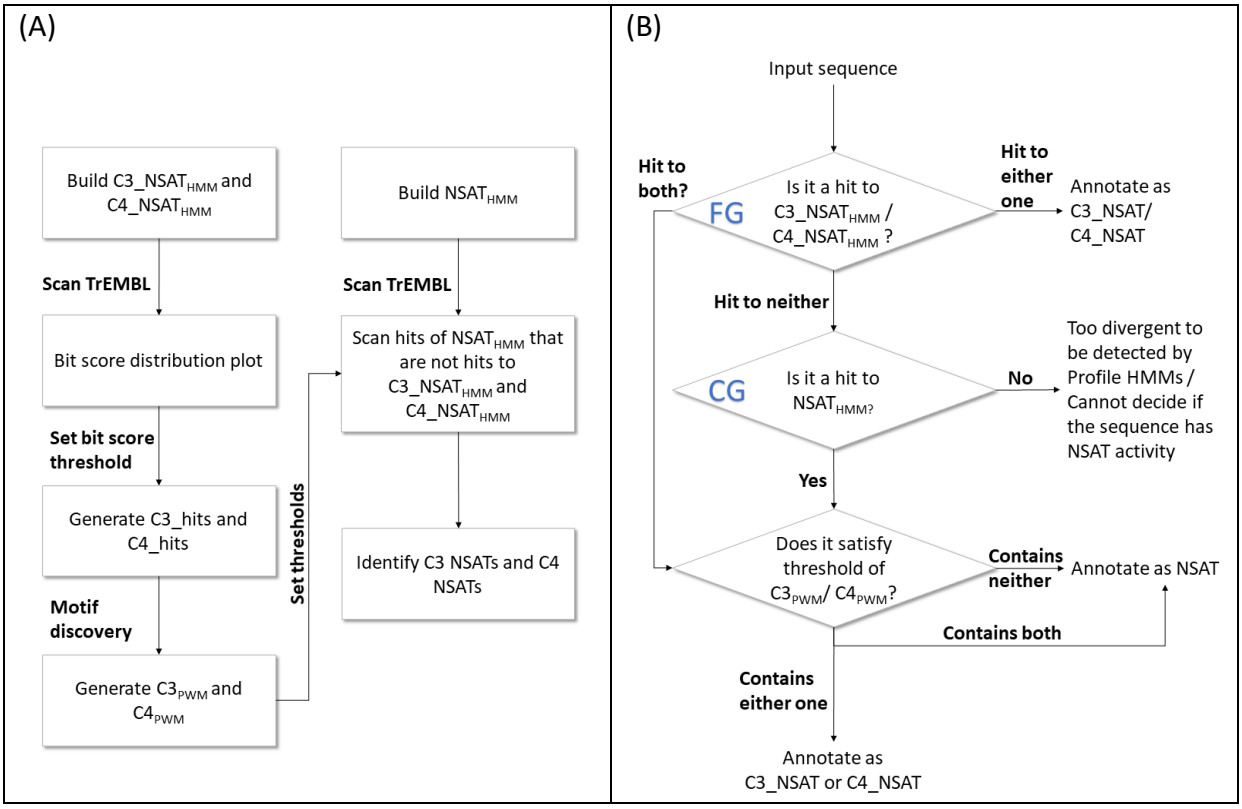


Figure 5: (A) Steps involved in identification of C3 and C4 NSATs from TrEMBL using profiles HMMs and PWMs. (B) Flowchart showing the steps involved in annotating an input sequence based on profile HMMs and PWMs. In steps marked FG and CG, annotations are Fine Grained, i.e., C3 NSATs an C4 NSATs, and Coarse Grained, i.e., NSATs (relative to each other; also see Figure 1).

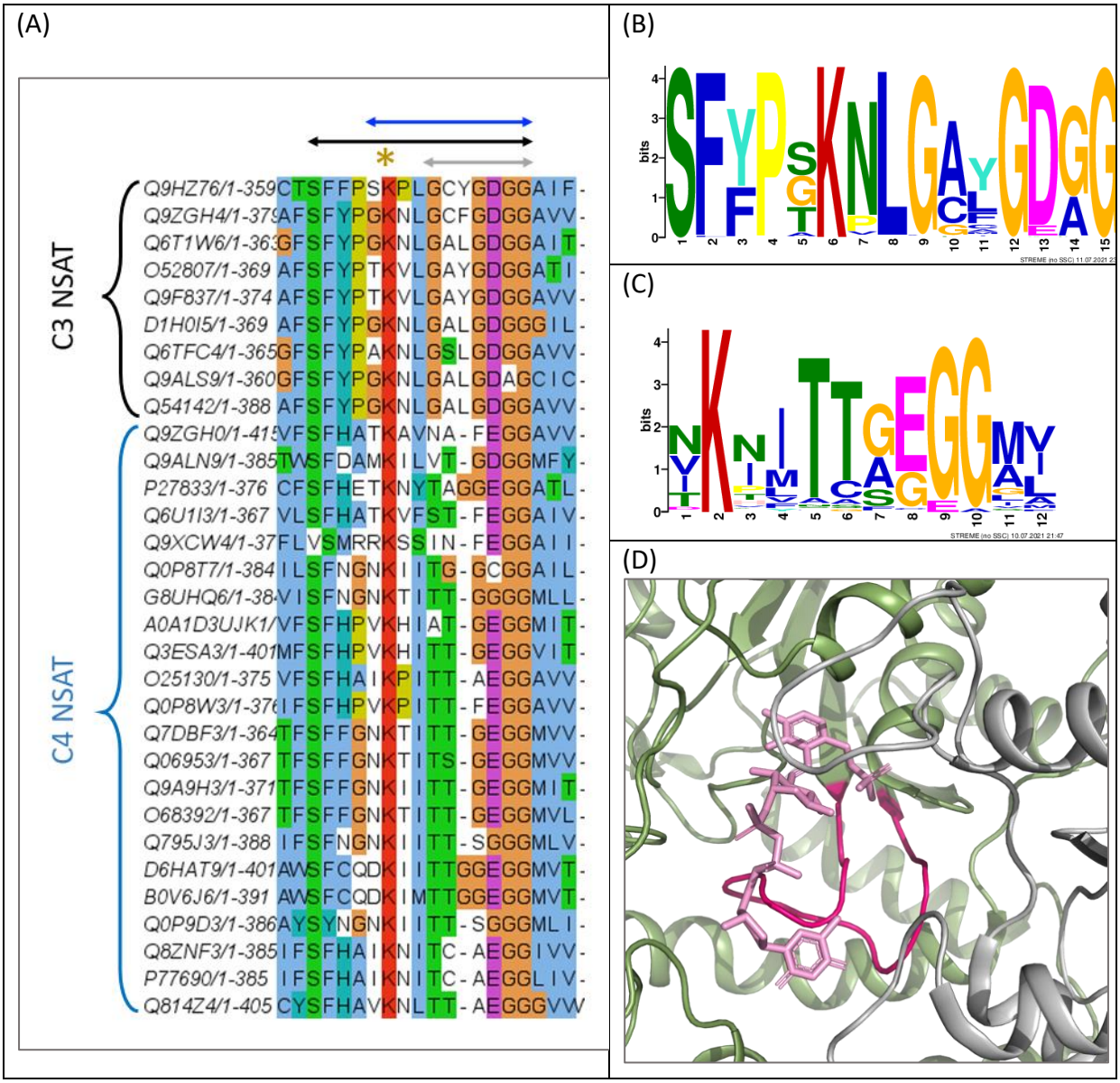


Figure 6: (A) A part of the multiple sequence alignment of C3 NSATs and C4 NSATs (full alignment is shown as Figure S4). The conserved lysine that serves the role of Schiff's base is marked with a brown asterisk. The visually distinct 7-residue stretch conserved in C3 NSATs, but not in C4 NSATs, is marked by the grey double-headed arrow. The 15-residue (in C3 NSATs) and 12-residue (in C4 NSATs) motifs identified by the STREME server are marked by black and blue double-headed arrows, respectively. Whether a 12-mer that excluded positions 1-3 of the 15-mer motif can be a potential C3_motif was also probed. It was found that the highest scoring 12-mers in six sequences from C3_hits were not from the expected position, but adjacent to it. Hence, this 12-mer motif was not considered for any further analysis.

Sequence logos of motifs obtained from STREME for C3_hits (A) and C4_hits (B). (D) The location of the motif in the 3D structure of a C3 NSAT viz., QdtB (PDB ID 3FRK) from *Thermoanaerobacterium thermosaccharolyticum*. The motif is in dark pink and the bound nucleotide sugar product is in light pink.

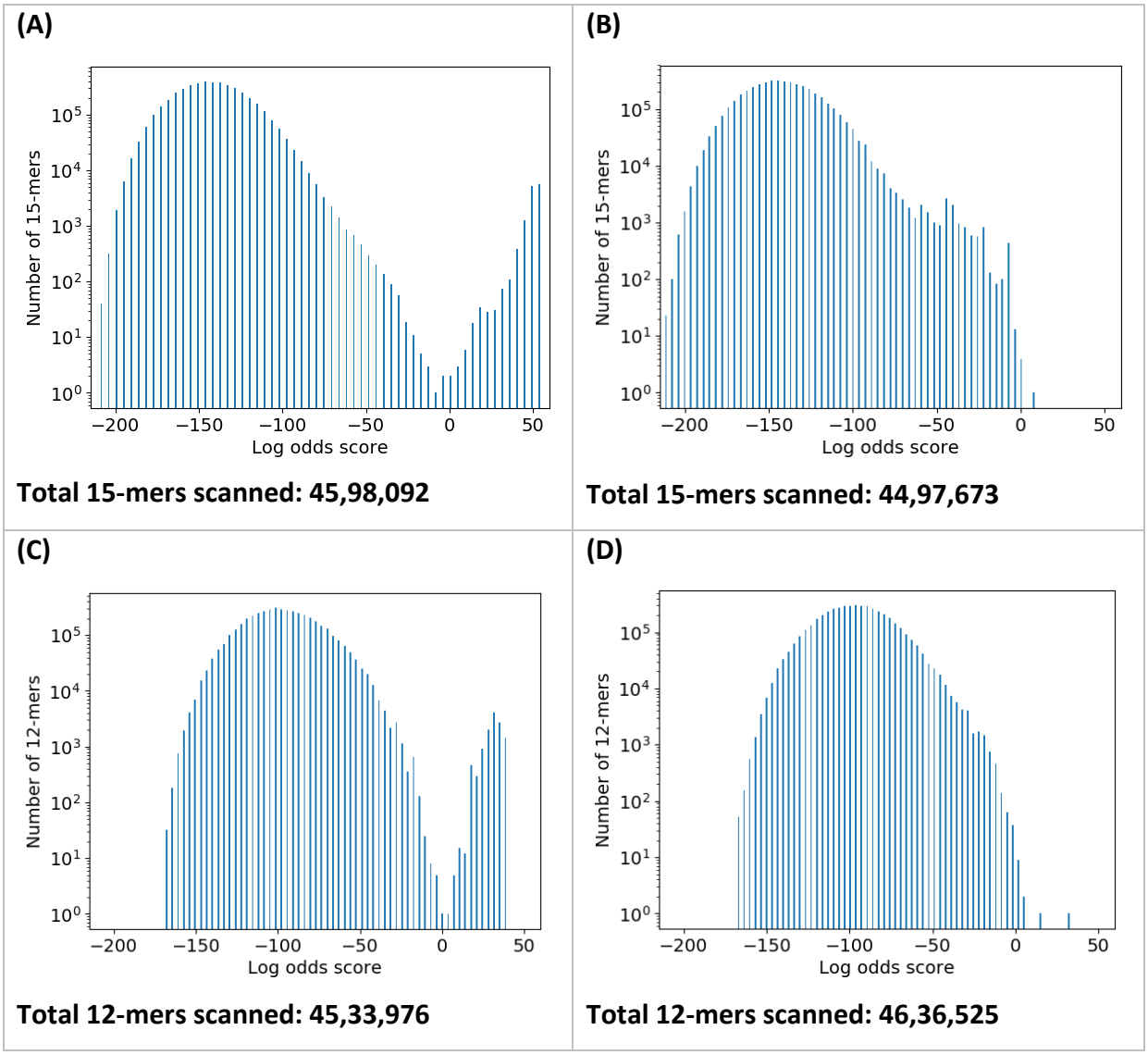


Figure 7: Distribution of log odds score of all 15-mers of C3_hits (A) and C4_hits (B) for the $C3_{PWM}$ and all 12-mers of C4_hits (C) and C3_hits (D) for the $C4_{PWM}$. Log-odds score is plotted in the X-axis and the number of 15-/12-mers (in log scale) are plotted in the Y-axis.