Clues to reaction specificity in PLP-dependent fold type I aminotransferases 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
Lab 402, Department of Biosciences and Bioengineering
IIT Bombay, Powai-400076
Mumbai
Email: jaya_srivastava@iitb.ac.in

Running title: Reaction specificity in PLP-dependent fold type I aminotransferases

Supplementary material Includes:

1. Supplementary file containing Tables S1-S3, Figures S1-S6
2. Supplementary data.xlsx

Abbreviations

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

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 (which will be henceforth referred to as aminotransferases) belong to fold type I and mediate the biosynthesis of several monosaccharides. They use diverse substrates but are highly selective to the C3 or C4 carbon to which amine group is transferred. Profile HMMs were able to identify aminotransferases but could not capture reaction specificity. A search for discriminating features led to the discovery of sequence motifs that are 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 aminotransferases. Inferences from this analysis set way for future experiments that can shed light on mechanisms of functional diversification in nucleotide sugar aminotransferases of fold type I.

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

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 (i) a few residues, such as cofactor binding lysine in PLP dependent enzymes[9], (ii) 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[13] (Figure 1) thereby reflecting catalytic strategies that have independently evolved among PLP dependent enzymes[12]. Nucleotide sugar aminotransferases and dehydratases belong to fold type I[14]. This group also contains enzymes that catalyze several other types of reactions[15] (Figure 1).

Aminotransferases and dehydratases take part in the biosynthesis of at least 19 monosaccharides (Figure S1) and secondary metabolites[16,17]. Aminotransferases convert keto-sugars to deoxy amino sugars. Dehydratases utilize similar substrates as aminotransferases but catalyse dehydration instead of transamination. For example, ColD (a dehydratase), utilizes the same substrate as Per (a C4 aminotransferase) but catalyzes

dehydration as opposed to transamination[18]. These catalytic differences among aminotransferases and dehydratases have emerged by keeping the overall structure intact (Figure 2). Even the extent of sequence similarity among functional variants of aminotransferases and dehydratases is so comparable (Figure S2) that it is not possible to distinguish between these two families solely based on pairwise sequence similarity. Residues which upon mutation could convert dehydratase to aminotransferase have been reported in literature[19,20]. However, features governing reaction specificity among aminotransferases remain unexplored.

Aminotransferases may 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 (Figure 3). Some of the experimentally characterized aminotransferases (Supplementary data.xslx, worksheet: Exp and reviewed sequences) 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]. Clearly, reaction specificity is determined by as yet unknown "minor" differences in primary structures. In this study, we have identified features that discriminate C3 aminotransferases from C4 aminotransferases using profile Hidden Markov Models (HMMs) and position weight matrices (PWMs).

2. METHODS

Databases, software and online servers: Protein sequences were obtained from UniProtKB (Release 2021_1)[25] and 3D structures from the Protein Data Bank[26]. 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[27]. Pairwise sequence comparisons were performed using BLASTp online server[28]. Pairwise identity matrix was obtained using MUSCLE webserver[29]. Multiple sequence alignments (MSAs) were done using the MAFFT command line application[30]. 3D structures were visualized, analyzed and compared using PyMol[31]. HMMER3[32] was used to build profile HMMs. Electron density maps of ligand structures were visualized using WinCOOT[33].

Datasets and profile HMMs: Experimentally characterized aminotransferases and dehydratases were obtained from literature; reviewed homologs were taken from SwissProt (Supplementary data.xslx, worksheet: Exp and reviewed sequences). Bit score thresholds for $NSAT_{HMM}$ and NSD_{HMM}, profile HMMs of aminotransferases and dehydratases, respectively, 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[6,34]. 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 scoring the experimentally characterized sequences of CATH superfamily 3.40.640.10 (fold type I PLP-dependent enzymes) [Supplementary data.xslx, worksheet: CATH-3.40.640.10]. 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_$NSAT_{HMM}$ (profile HMMs for C3 and C4 aminotransferases, respectively) since a majority of annotations in TrEMBL just mention aminotransferase without specifying if the protein is a C3 aminotransferase or C4 aminotransferase. 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[35] was used to identify sequence motifs enriched in a sequence family with reference to a set of "control" sequences. By subjecting TrEMBL hits of C3_$NSAT_{HMM}$ and C4_$NSAT_{HMM}$ to a redundancy criterion of 100% identity, a total of 12,811 and 12,101 sequences were obtained and designated as C3_hits and C4_hits, respectively. The dataset C4_hits 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[36]. Experimentally characterized sequences from CATH superfamily 3.40.640.10 (Supplementary data.xslx, worksheet: CATH-3.40.640.10) with >=80% identity were removed, and the resulting 208 enzymes, including PglE from *Camplyobacter jejuni* (PDB ID: 4ZTC)[37], a C4 aminotransferase that is structurally identical to other C4 aminotransferases 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 aminotransferases using profile HMMs: Scoring 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.xslx, worksheet: common_hits). Among sequences scoring below 400 against both profiles, 32,344 were hits to NSAT$_{HMM}$. These sequences are predicted to be aminotransferases (because they are hits of NSAT$_{HMM}$) but we could not resolve if these are C3 or C4 aminotransferases, or neither of these two. Similarly, we could not resolve if 21 common hits are C3 or C4 aminotransferases or have both activities. Thus, our inability to set a bit score threshold that is sensitive enough to identify C3 and C4 aminotransferases without compromising specificity meant that an alternative approach is required to discriminate C3 and C4 aminotransferases from each other.

Identifying motifs that discriminate C3 aminotransferases and C4 aminotransferases from each other: C3_hits and C4_hits were used to find motifs enriched in C3 aminotransferases

in comparison to C4 aminotransferases and vice versa using STREME[a]. The most conserved motif among C3_hits is a 15-mer found in 99.9% of input sequences (Figure 6A,6B). The most conserved motif among C4 aminotransferases is a 12-mer located at the same site as the conserved motif in C3_hits (Figure 6A,6C). This motif is found in all but one C4_exp sequences. Visual inspection of the sequence logos confirmed the distinctiveness of these motifs (Figure 6A). Specifically, the conserved 15-mer of C3_hits contain a glycine residue at position 9 (Figure 6B), whose role has not been investigated so far. Both motifs map to the active site of aminotransferases (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 region that best matches the 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).

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 aminotransferase 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

[a] 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

aminotransferase 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 aminotransferases and C4 aminotransferases that could not be identified using profile HMMs.

Resolving common hits: As mentioned earlier, 21 sequences are hits to both $C3_NSAT_{HMM}$ and $C4_NSAT_{HMM}$ (Supplementary data.xslx, worksheet: common_hits). They were scanned using $C3_{PWM}$ and $C4_{PWM}$. 19 sequences are distinctly identified as C3 aminotransferase or C4 aminotransferase 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 aminotransferases and C4 aminotransferases 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 scored 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 aminotransferases and C4 aminotransferases (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 aminotransferases and 5,887 as C3 aminotransferases. It can be observed that C3 aminotransferases score less than C4 aminotransferases against $C4_NSAT_{HMM}$; however, the converse is not true i.e., C4 aminotransferases do not necessarily score less than C3 aminotransferases 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

aminotransferase activity and acquired novel function(s) due to sequence changes that are too subtle to be detected by HMMs and PWMs.

Validating the prediction pipeline with putative aminotransferases of known biosynthesis pathways: We identified incompletely characterized sugar biosynthesis pathways from literature that involve putative aminotransferases to validate our two-step prediction pipeline. Seven such aminotransferases were identified (Supplementary data.xslx, worksheet: Test dataset). Four among these were identified as C3 aminotransferases using profile HMMs. The rest, OleNI, KijD7, and EryCIV, were identified as C4 aminotransferases using profile HMM and PWM-based search. OleNI was earlier proposed to act as a 3,4-dehydratase in desosamine biosynthesis[39]. However, the pathway for TDP-desosamine has been revised using homologous genes from *Streptomyces venezuelae*[40] and *Saccharopolyspora erythraea*[41] and supports our prediction of OleNI as a C4 aminotransferase.

Additional factors governing product identity - clues from 3D structures: We superposed the 3D structures of C3 and C4 aminotransferases bound to substrates in the form of external aldimine (Table 2). The backbone conformation is highly similar in these six proteins (Figure 2A). In addition, the overall 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 S4A). However, PLP forms aldimine with C3 keto or C4 keto group and this difference is accommodated by conformation change around the pyrophosphate group (Figure S4A). The conformation of ligands is identical in C3 aminotransferases (Figure S4B); in the three C4 aminotransferases, the pyranose ring either flips or shifts (Figure S4C, S4D). 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 aminotransferases, 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 ~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[42]. 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.

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 aminotransferases, dehydratases, non-nucleotide sugar aminotransferases and AHBA synthase[43]. We obtained a similar sugar AT family of 22 enzymes from a sequence based phylogenetic analysis (Figure S6). 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 between C3 and C4 aminotransferases and between aminotransferases and dehydratases?

Aside from sequence and structural similarity among aminotransferases, conserved PLP and nucleotide group bound conformations suggest that differences are confined to regions that interact with pyranose ring (Figure S4). Using a combination of profile HMMs and motif based PWMs, we located a sequence motif in the pyranose binding site that differs among C3 and C4 aminotransferases (Figure 6). Structural comparisons led to the identification of a tyrosine residue that forms H-bond with pyranose ring and is conserved among C3 aminotransferases but not in C4 aminotransferases (Table 2). This suggests that absence of stabilizing interactions may lead to flipping or shifting of pyranose ring to orient C4 keto group for amination. H-bonding residues in that position are conserved among C3_hits (as tyrosine) but only in ~40% of the C4_hits. Thus, tyrosine can be one of the factors that determine the shift or flip observed in pyranose ring.

Other enzymes in the 'sugar AT family' contain dehydratases that utilize nucleotide linked sugars but catalyse dehydration instead. 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[18,44] but by two enzymes in the pathway for the biosynthesis of CDP-linked 3,6-dideoxy sugars[19]. A hallmark of dehydratases 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 [20]. Additional residues, i.e., Asp194, Tyr217 and Phe345 in *Yersinia pseudotuberculosis* E1 dehydrase and Ser187 in *Escherichia coli* ColD, have been shown to convert a dehydratase to an aminotransferase [19,20]. However, among experimentally characterized dehydratases (Supplementary data.xslx, worksheet: Exp and reviewed sequences), Ser187 is not conserved, D194 is also found substituted with glutamate which is found among several C4 aminotransferases, and Y217 is conserved among C3 aminotransferases as well. We sought to analyse the extent of conservation in these 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) among NSD_hits (a dataset of 4018 TrEMBL entries that satisfy the bit score threshold of NSD_{HMM}), C3_hits and C4_hits (Table S3). Observations from this analysis suggest that not all these sites are strictly conserved among dehydratases. 'Sugar AT family' also contains aminotransferases that utilize sugars not linked to nucleotides. NtdA, a 3-oxo-glucose-6-phosphate aminotransferase[45], and KdnA, a 8-amino-3,8-dideoxy-alpha-D-*manno*-octulosonate aminotransferase[46] 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[45]. In RbmB, a 2-deoxy-*scyllo*-inosose aminotransferase[47], 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 fold space of polypeptides is finite[48–52]. Consequently, the same fold can accommodate multiple functions using diverse mechanisms such as point mutations, indels, recruitment of additional domains, oligomerization, etc.[53]. Several factors such as gene loss, enzyme promiscuity, biological networks, etc., are proposed to drive neo- and subfunctionalization[54–56]. Aminotransferases and dehydratases are examples of functional families wherein a few mutations lead to alterations in reaction specificity[19,20]. In the absence of adequate amount of experimental data that capture sequence changes to functional changes at the level of substrate and reaction specificity, 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 and C4 aminotransferases 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 aminotransferases, 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[57]. Minor alterations in the active site leading to functional diversity are being routinely engineered for biotechnological applications[58,59]. With experimental validation, features highlighted in this study that contribute to reaction specificities in aminotransferases and its homologs may be useful in engineering efficient enzymes for the biosynthesis of natural products.

Data availability

PDB codes used in this study can be accessed from RCSB Protein Data Bank.

Conflicts of Interests

None declared.

Author Contributions

PVB conceptualized this study; JS performed analysis; JS and PVB co-wrote 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. Gabius HJ, Roth J. An introduction to the sugar code. *Histochem Cell Biol*. 2017;147(2):111-117. doi:10.1007/s00418-016-1521-9

2. Mostowy RJ, Holt KE. Diversity-Generating Machines: Genetics of Bacterial Sugar-Coating. *Trends Microbiol*. 2018;26(12):1008-1021. doi:10.1016/j.tim.2018.06.006

3. Thibodeaux CJ, Melançon CE, Liu H wen. Natural-product sugar biosynthesis and enzymatic glycodiversification. *Angew Chem Int Ed Engl*. 2008;47(51):9814-9859. doi:10.1002/anie.200801204

4. Gabius HJ. The sugar code: Why glycans are so important. *Biosystems*. 2018;164:102-111. doi:10.1016/j.biosystems.2017.07.003

5. Seeberger PH. Monosaccharide Diversity. In: Varki A, Cummings RD, Esko JD, et al., eds. *Essentials of Glycobiology*. 3rd ed. Cold Spring Harbor Laboratory Press; 2015. Accessed December 8, 2020. http://www.ncbi.nlm.nih.gov/books/NBK453086/

6. Srivastava J, Sunthar P, Balaji PV. The glycan alphabet is not universal: a hypothesis. *Microb Genom*. 2020;6(11). doi:10.1099/mgen.0.000452

7. Thoden JB, Holden HM. Structural and functional studies of WlbA: A dehydrogenase involved in the biosynthesis of 2,3-diacetamido-2,3-dideoxy-D-mannuronic acid. *Biochemistry*. 2010;49(36):7939-7948. doi:10.1021/bi101103s

8. Taberman H, Parkkinen T, Rouvinen J. Structural and functional features of the NAD(P) dependent Gfo/Idh/MocA protein family oxidoreductases. *Protein Sci*. 2016;25(4):778-786. doi:10.1002/pro.2877

9. Oliveira EF, Cerqueira NMFSA, Fernandes PA, Ramos MJ. Mechanism of formation of the internal aldimine in pyridoxal 5'-phosphate-dependent enzymes. *J Am Chem Soc*. 2011;133(39):15496-15505. doi:10.1021/ja204229m

10. Vey JL, Drennan CL. Structural insights into radical generation by the radical SAM superfamily. *Chem Rev*. 2011;111(4):2487-2506. doi:10.1021/cr9002616

11. Da Costa M, Gevaert O, Van Overtveldt S, et al. Structure-function relationships in NDP-sugar active SDR enzymes: Fingerprints for functional annotation and enzyme engineering. *Biotechnol Adv*. 2021;48:107705. doi:10.1016/j.biotechadv.2021.107705

12. Schneider G, Käck H, Lindqvist Y. The manifold of vitamin B6 dependent enzymes. *Structure*. 2000;8(1):R1-6. doi:10.1016/s0969-2126(00)00085-x

13. B6db home : Main Menu. Accessed July 13, 2021. http://bioinformatics.unipr.it/cgi-bin/bioinformatics/B6db/home.pl

14. Fesko K, Suplatov D, Švedas V. 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*. 2018;8(6):1013-1028. doi:10.1002/2211-5463.12441

15. Catazaro J, Caprez A, Guru A, Swanson D, Powers R. Functional evolution of PLP-dependent enzymes based on active-site structural similarities. *Proteins*. 2014;82(10):2597-2608. doi:10.1002/prot.24624

16. Romo AJ, Liu H wen. Mechanisms and structures of vitamin B(6)-dependent enzymes involved in deoxy sugar biosynthesis. *Biochim Biophys Acta*. 2011;1814(11):1534-1547. doi:10.1016/j.bbapap.2011.02.003

17. Timmons SC, Thorson JS. Increasing carbohydrate diversity via amine oxidation: aminosugar, hydroxyaminosugar, nitrososugar, and nitrosugar biosynthesis in bacteria. *Curr Opin Chem Biol*. 2008;12(3):297-305. doi:10.1016/j.cbpa.2008.03.017

18. Alam J, Beyer N, Liu H wen. 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*. 2004;43(51):16450-16460. doi:10.1021/bi0483763

19. Smith P, Szu PH, Bui C, Liu H wen, Tsai SC. Structure and Mutagenic Conversion of E1 Dehydrase: at the Crossroads of Dehydration, Aminotransfer and Racemization,. *Biochemistry*. 2008;47(24):6329-6341. doi:10.1021/bi702449p

20. Cook PD, Kubiak RL, Toomey DP, Holden HM. Two site-directed mutations are required for the conversion of a sugar dehydratase into an aminotransferase. *Biochemistry*. 2009;48(23):5246-5253. doi:10.1021/bi9005545

21. Larkin A, Imperiali B. 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*. 2009;48(23):5446-5455. doi:10.1021/bi900186u

22. Useglio M, Peirú S, Rodríguez E, Labadie GR, Carney JR, Gramajo H. TDP-L-megosamine biosynthesis pathway elucidation and megalomicin a production in Escherichia coli. *Appl Environ Microbiol*. 2010;76(12):3869-3877. doi:10.1128/AEM.03083-09

23. Hwang S, Li Z, Bar-Peled Y, Aronov A, Ericson J, Bar-Peled M. 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*. 2014;289(51):35620-35632. doi:10.1074/jbc.M114.614917

24. Thoden JB, Schäffer C, Messner P, Holden HM. Structural analysis of QdtB, an aminotransferase required for the biosynthesis of dTDP-3-acetamido-3,6-dideoxy-alpha-D-glucose. *Biochemistry*. 2009;48(7):1553-1561. doi:10.1021/bi8022015

25. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*. 2019;47(D1):D506-D515. doi:10.1093/nar/gky1049

26. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. *Nucleic Acids Res*. 2000;28(1):235-242. doi:10.1093/nar/28.1.235

27. Sillitoe I, Bordin N, Dawson N, et al. CATH: increased structural coverage of functional space. *Nucleic Acids Res*. 2021;49(D1):D266-D273. doi:10.1093/nar/gkaa1079

28. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403-410. doi:10.1016/S0022-2836(05)80360-2

29. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32(5):1792-1797. doi:10.1093/nar/gkh340

30. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*. 2013;30(4):772-780. doi:10.1093/molbev/mst010

31. PyMOL | pymol.org. Accessed May 18, 2021. https://pymol.org/2/

32. Eddy SR. A new generation of homology search tools based on probabilistic inference. *Genome Inform*. 2009;23(1):205-211.

33. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr*. 2004;60(Pt 12 Pt 1):2126-2132. doi:10.1107/S0907444904019158

34. Srivastava J, Hembrom R, Kumawat A, Balaji PV. Searching sequence databases for functional homologs using profile HMMs: how to set bit score thresholds? *bioRxiv*. Published online June 25, 2021:2021.06.24.449764. doi:10.1101/2021.06.24.449764

35. Bailey TL. STREME: Accurate and versatile sequence motif discovery. *Bioinformatics*. Published online March 24, 2021. doi:10.1093/bioinformatics/btab203

36. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol*. 2016;33(7):1870-1874. doi:10.1093/molbev/msw054

37. Riegert AS, Young NM, Watson DC, Thoden JB, Holden HM. Structure of the external aldimine form of PglE, an aminotransferase required for N,N'-diacetylbacillosamine biosynthesis. *Protein Sci*. 2015;24(10):1609-1616. doi:10.1002/pro.2745

38. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*. 2021;49(W1):W293-W296. doi:10.1093/nar/gkab301

39. Aguirrezabalaga I, Olano C, Allende N, et al. Identification and expression of genes involved in biosynthesis of L-oleandrose and its intermediate L-olivose in the oleandomycin producer Streptomyces antibioticus. *Antimicrob Agents Chemother*. 2000;44(5):1266-1275. doi:10.1128/AAC.44.5.1266-1275.2000

40. Zhao L, Borisova S, Yeung SM, Liu H. Study of C-4 deoxygenation in the biosynthesis of desosamine: evidence implicating a novel mechanism. *J Am Chem Soc*. 2001;123(32):7909-7910. doi:10.1021/ja010587x

41. Jiang M, Zhang H, Park SH, Li Y, Pfeifer BA. Deoxysugar pathway interchange for erythromycin analogues heterologously produced through Escherichia coli. *Metab Eng*. 2013;20:92-100. doi:10.1016/j.ymben.2013.09.005

42. Burgie ES, Holden HM. Molecular architecture of DesI: a key enzyme in the biosynthesis of desosamine. *Biochemistry*. 2007;46(31):8999-9006. doi:10.1021/bi700751d

43. Itoh Y, Bröcker MJ, Sekine S ichi, Söll D, Yokoyama S. Dimer-dimer interaction of the bacterial selenocysteine synthase SelA promotes functional active-site formation and catalytic specificity. *J Mol Biol*. 2014;426(8):1723-1735. doi:10.1016/j.jmb.2014.01.003

44. Hong L, Zhao Z, Melançon CE, Zhang H, Liu H wen. In vitro characterization of the enzymes involved in TDP-D-forosamine biosynthesis in the spinosyn pathway of Saccharopolyspora spinosa. *J Am Chem Soc*. 2008;130(14):4954-4967. doi:10.1021/ja0771383

45. van Straaten KE, Ko JB, Jagdhane R, Anjum S, Palmer DRJ, Sanders DAR. 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*. 2013;288(47):34121-34130. doi:10.1074/jbc.M113.500637

46. Zachman-Brockmeyer TR, Thoden JB, Holden HM. Structures of KdnB and KdnA from Shewanella oneidensis: Key Enzymes in the Formation of 8-Amino-3,8-Dideoxy-d-Manno-Octulosonic Acid. *Biochemistry*. 2016;55(32):4485-4494. doi:10.1021/acs.biochem.6b00439

47. Zachman-Brockmeyer TR, Thoden JB, Holden HM. The structure of RbmB from Streptomyces ribosidificus, an aminotransferase involved in the biosynthesis of ribostamycin. *Protein Sci*. 2017;26(9):1886-1892. doi:10.1002/pro.3221

48. Schaeffer RD, Daggett V. Protein folds and protein folding. *Protein Eng Des Sel*. 2011;24(1-2):11-19. doi:10.1093/protein/gzq096

49. Liu X, Lv B, Guo W. The size distribution of protein families within different types of folds. *Biochem Biophys Res Commun*. 2011;406(2):218-222. doi:10.1016/j.bbrc.2011.02.020

50. Govindarajan S, Recabarren R, Goldstein RA. Estimating the total number of protein folds. *Proteins*. 1999;35(4):408-414.

51. Kolodny R, Pereyaslavets L, Samson AO, Levitt M. On the universe of protein folds. *Annu Rev Biophys*. 2013;42:559-582. doi:10.1146/annurev-biophys-083012-130432

52. Yang G, Miton CM, Tokuriki N. A mechanistic view of enzyme evolution. *Protein Sci*. 2020;29(8):1724-1747. doi:10.1002/pro.3901

53. Copley SD. Evolution of new enzymes by gene duplication and divergence. *FEBS J*. 2020;287(7):1262-1283. doi:10.1111/febs.15299

54. Juárez-Vázquez AL, Edirisinghe JN, Verduzco-Castro EA, et al. Evolution of substrate specificity in a retained enzyme driven by gene loss. *Elife*. 2017;6:e22679. doi:10.7554/eLife.22679

55. Glasner ME, Truong DP, Morse BC. How enzyme promiscuity and horizontal gene transfer contribute to metabolic innovation. *FEBS J*. 2020;287(7):1323-1342. doi:10.1111/febs.15185

56. Roth C, Rastogi S, Arvestad L, et al. Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. *J Exp Zool B Mol Dev Evol*. 2007;308(1):58-73. doi:10.1002/jez.b.21124

57. F. Rocha J, F. Pina A, F. Sousa S, A. Cerqueira NMFS. PLP-dependent enzymes as important biocatalysts for the pharmaceutical, chemical and food industries: a structural and mechanistic perspective. *Catalysis Science & Technology*. 2019;9(18):4864-4876. doi:10.1039/C9CY01210A

58. Fernandez FJ, de Vries D, Peña-Soler E, et al. Structure and mechanism of a cysteine sulfinate desulfinase engineered on the aspartate aminotransferase scaffold. *Biochim Biophys Acta*. 2012;1824(2):339-349. doi:10.1016/j.bbapap.2011.10.016

59. Steffen-Munsberg F, Vickers C, Kohls H, et al. Bioinformatic analysis of a PLP-dependent enzyme superfamily suitable for biocatalytic applications. *Biotechnol Adv*. 2015;33(5):566-604. doi:10.1016/j.biotechadv.2014.12.012

60. Ishiyama N, Creuzenet C, Miller WL, et al. Structural studies of FlaA1 from Helicobacter pylori reveal the mechanism for inverting 4,6-dehydratase activity. *J Biol Chem*. 2006;281(34):24489-24495. doi:10.1074/jbc.M602393200

61. Holliday GL, Akiva E, Meng EC, et al. Atlas of the Radical SAM Superfamily: Divergent Evolution of Function Using a "Plug and Play" Domain. *Methods Enzymol*. 2018;606:1-71. doi:10.1016/bs.mie.2018.06.004

62. Dow GT, Gilbert M, Thoden JB, Holden HM. Structural investigation on WlaRG from Campylobacter jejuni: A sugar aminotransferase. *Protein Sci*. 2017;26(3):586-599. doi:10.1002/pro.3109

63. Burgie ES, Thoden JB, Holden HM. Molecular architecture of DesV from Streptomyces venezuelae: a PLP-dependent transaminase involved in the biosynthesis of the unusual sugar desosamine. *Protein Sci*. 2007;16(5):887-896. doi:10.1110/ps.062711007

64. Wang F, Singh S, Xu W, et al. Structural Basis for the Stereochemical Control of Amine Installation in Nucleotide Sugar Aminotransferases. *ACS Chem Biol*. 2015;10(9):2048-2056. doi:10.1021/acschembio.5b00244

65. Cook PD, Carney AE, Holden HM. Accommodation of GDP-linked sugars in the active site of GDP-perosamine synthase. *Biochemistry*. 2008;47(40):10685-10693. doi:10.1021/bi801309q

66. Schoenhofen IC, Lunin VV, Julien JP, et al. Structural and functional characterization of PseC, an aminotransferase involved in the biosynthesis of pseudaminic acid, an essential flagellar modification in Helicobacter pylori. *J Biol Chem*. 2006;281(13):8907-8916. doi:10.1074/jbc.M512987200

67. Noland BW, Newman JM, Hendle J, et al. Structural studies of Salmonella typhimurium ArnB (PmrH) aminotransferase: a 4-amino-4-deoxy-L-arabinose lipopolysaccharide-modifying enzyme. *Structure*. 2002;10(11):1569-1580. doi:10.1016/s0969-2126(02)00879-1

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

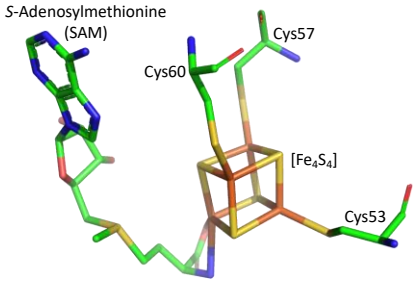
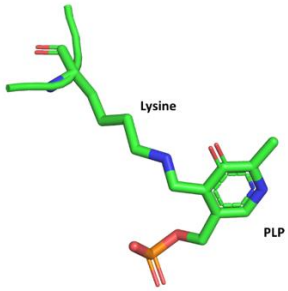
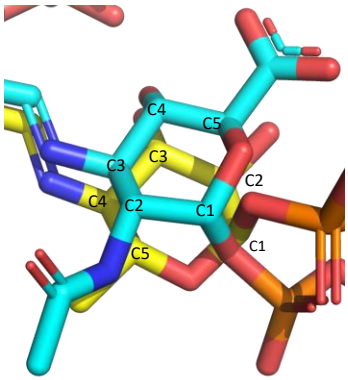
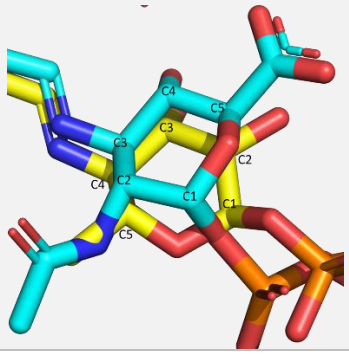
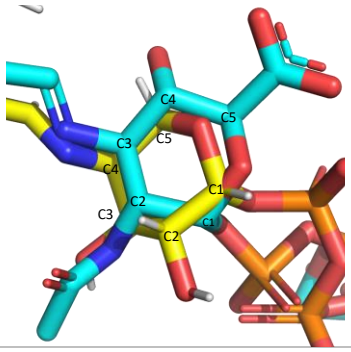
Conserved feature	Representative 3D rendering
FOLD: SDR superfamily Representative example from [60] Rendering: PDB ID 2GN4. Region highlighted in pink shows the core Rossmann fold domain of SDR superfamily **Associated functions**: Dehydratase, dehydrogenase, enolase, epimerase, isomerase, reductase, etc.	
SEQUENCE MOTIF: Cys-X_3-Cys-X_2-Cys Representative example from [61] Rendering: PDB id 1R30 **Associated functions**: Dehydrogenase, lyase, methyltransferase, etc.	
RESIDUE: Lysine in vitamin B_6 dependent enzymes Representative example from [62] Rendering: PDB ID 3NYU **Associated functions**: Decarboxylase, dehydrase, ligase, lyase, mutase, phosphorylase, transaminase, etc.	

Table 2: Aminotransferases 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 S5, 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.xslx, 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 S4.
[†] Tyr309 of WbpE is conserved in 99.6% of C3_hits. The homologous residue is a tyrosine in only 40% of C4_hits.

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, aminotransferases are further classified as C3 aminotransferases and C4 aminotransferases. These, in turn, can install amine either axially or equatorially (assuming $^4C_1$ conformation) [Table S1, S2].

Figure 2A: Structural superimposition of four C3 aminotransferases and six C4 aminotransferases. PDB ids, substrates and orientation of the amino group in the product are given below.

C3 aminotransferases		
3NYU[62]	UDP-2-acetamido-2-deoxy-3-oxo-glucuronate	Equatorial ($^4C_1$)
5U21[62]	TDP-3-keto-6-deoxy-galactose/glucose	Equatorial ($^4C_1$)
3FRK[24]	TDP-3-keto-6-deoxy-glucose	Equatorial ($^4C_1$)
2OGA[63]	TDP-3-keto-4,6-dideoxy-D-*erythro*-hexopyranose	Equatorial ($^4C_1$)
C4 aminotransferases		
4ZAH[64]	TDP-4-keto-6-deoxy-glucose	Axial ($^4C_1$)
2PO3[42]	TDP-4-keto-6-deoxy-glucose	Equatorial ($^4C_1$)
3DR4[65]	GDP-4-keto-6-deoxy-mannose	Equatorial ($^4C_1$)
2FNU[66]	UDP-2-acetamido-2,6-dideoxy-beta-L-*arabino*-4-hexulose	Equatorial ($^4C_1$)
1MDO[67]	UDP-beta-L-*threo*-pentapyranos-4-ulose	Equatorial ($^4C_1$)
4ZTC[37]	UDP-4-keto-6-deoxy-GlcNAc	Equatorial ($^4C_1$)

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

Figure 3: Composite figures showing the reactants and products of C3 aminotransferases (A) and C4 aminotransferases (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. Ax (axial) and Eq (equatorial) are included in parenthesis merely to bring out the similarities in the orientations of the substituents on the pyranose rings in the substrates of various enzymes; here, the $^4C_1$ conformation is assumed for the pyranose ring. Schematic structures of reactants and products are given in Tables S1, S2.

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 aminotransferases 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 aminotransferases 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 aminotransferases an C4 aminotransferases, and Coarse Grained, i.e., aminotransferases (relative to each other; also see Figure 1).

Figure 6: (A) A part of the multiple sequence alignment of C3 and C4 aminotransferases. The conserved lysine that serves the role of Schiff's base is marked with a brown asterisk. The 15-residue (in C3 aminotransferases) and 12-residue (in C4 aminotransferases) motifs identified by the STREME server are marked by black and blue double-headed arrows, respectively. Whether a 12-mer corresponding to the 12-mer of C4 aminotransferases 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 **(B)** and C4_hits **(C)**. **(D)** The location of the motif in the 3D structure of a C3 aminotransferase viz., QdtB from *Thermoanaerobacterium thermosaccharolyticum* (PDB ID 3FRK)[24]. 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.

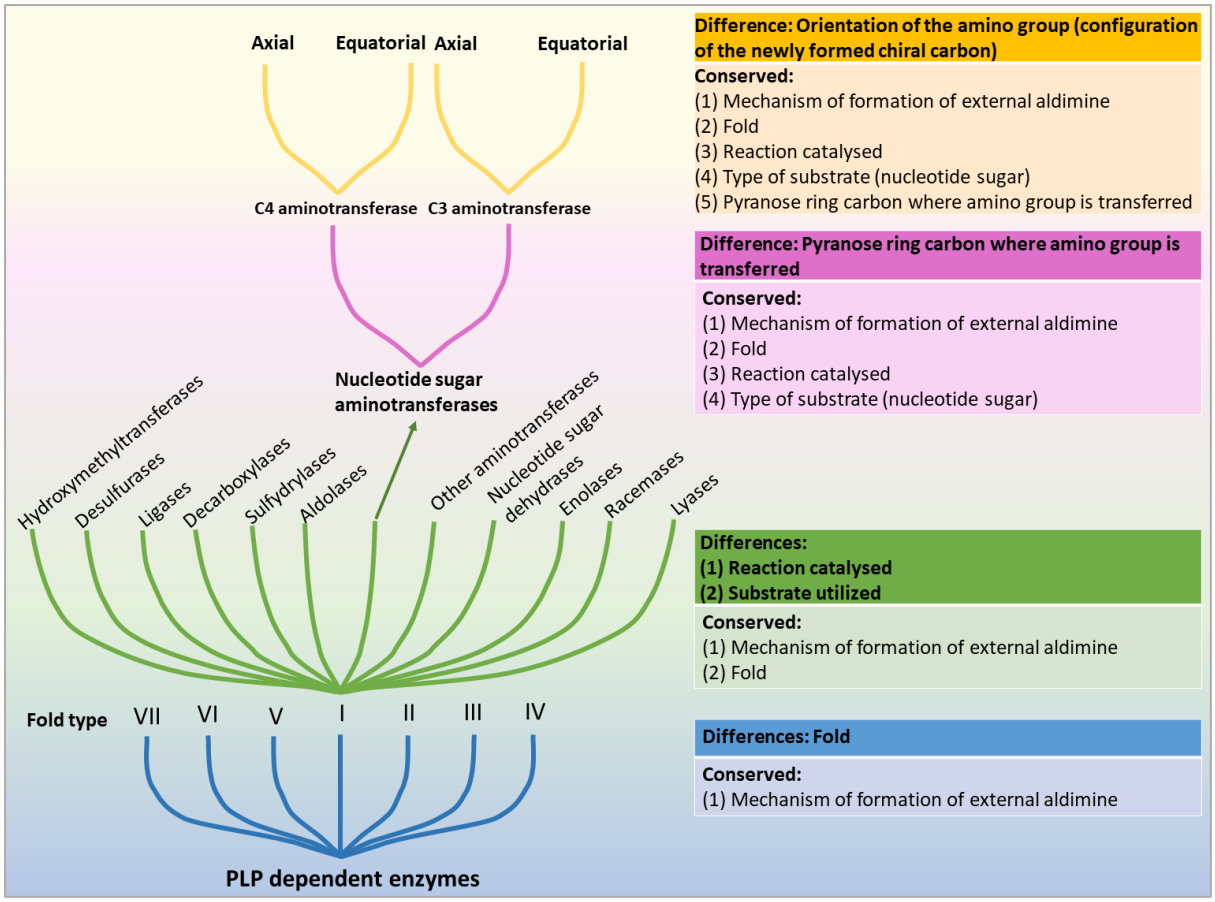
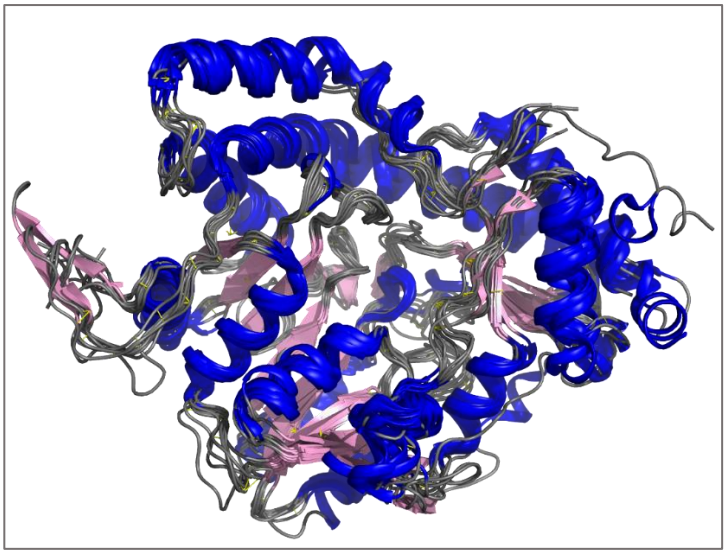

Axial
Equatorial
Axial
Equatorial
C4 aminotransferase
C3 aminotransferase
Nucleotide sugar aminotransferases
Hydroxymethyltransferases
Desulfurases
Ligases
Decarboxylases
Sulfydrylases
Aldolases
Other aminotransferases
Nucleotide sugar dehydrases
Enolases
Racemases
Lyases
Fold type
VII
VI
V
I
II
III
IV
PLP dependent enzymes
Difference: Orientation of the amino group (configuration of the newly formed chiral carbon)
Conserved:
(1) Mechanism of formation of external aldimine
(2) Fold
(3) Reaction catalysed
(4) Type of substrate (nucleotide sugar)
(5) Pyranose ring carbon where amino group is transferred
Difference: Pyranose ring carbon where amino group is transferred
Conserved:
(1) Mechanism of formation of external aldimine
(2) Fold
(3) Reaction catalysed
(4) Type of substrate (nucleotide sugar)
Differences:
(1) Reaction catalysed
(2) Substrate utilized
Conserved:
(1) Mechanism of formation of external aldimine
(2) Fold
Differences: Fold
Conserved:
(1) Mechanism of formation of external aldimine

Figure 1

(A)



(B)

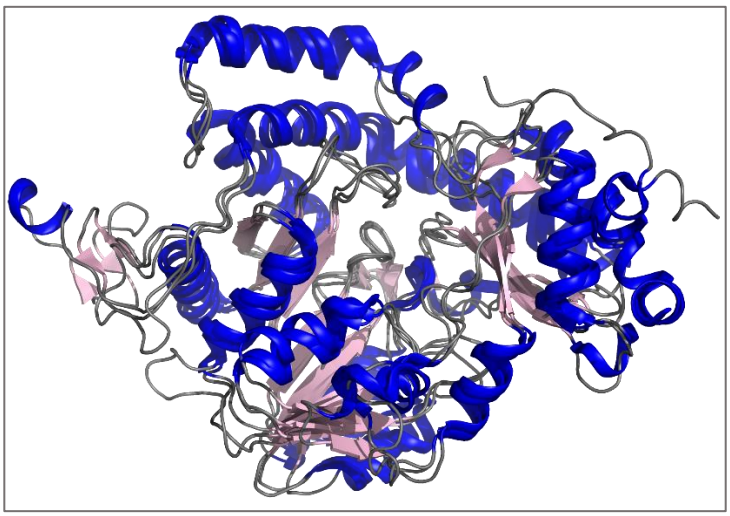
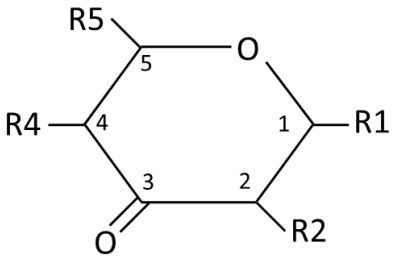


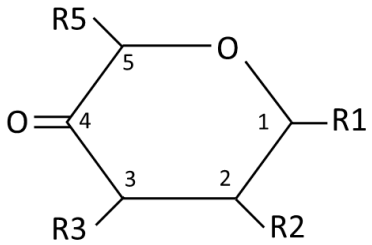
Figure 2

(A)



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

(B)



Gene name	R1	R2	R3	R5	C4-NH_2
DesI, VioA	TDP (Ax)	OH (Eq)	OH (Eq)	CH_3 (Eq)	(Eq)
WecE	TDP (Ax)	OH (Eq)	OH (Eq)	CH_3 (Eq)	(Ax)
LegC	GDP (Ax)	NAc (Eq)	OH (Eq)	CH_3 (Eq)	(Eq)
WeeJ, PglC, PglE, EpsN	UDP (Ax)	NAc (Eq)	OH (Eq)	CH_3 (Eq)	(Eq)
Pat	UDP (Ax)	NAc (Eq)	OH (Eq)	CH_3 (Eq)	(Ax)
PseC, Pam	UDP (Eq)	NAc (Ax)	OH (Ax)	CH_3 (Eq)	(Eq)
Per, PerA, RfbE	GDP (Ax)	OH (Ax)	OH (Eq)	CH_3 (Eq)	(Eq)
ArnB	UDP (Ax)	OH (Eq)	OH (Eq)	H	(Ax)
SpnR	TDP (Ax)	H	H	CH_3 (Eq)	(Eq)

Figure 3

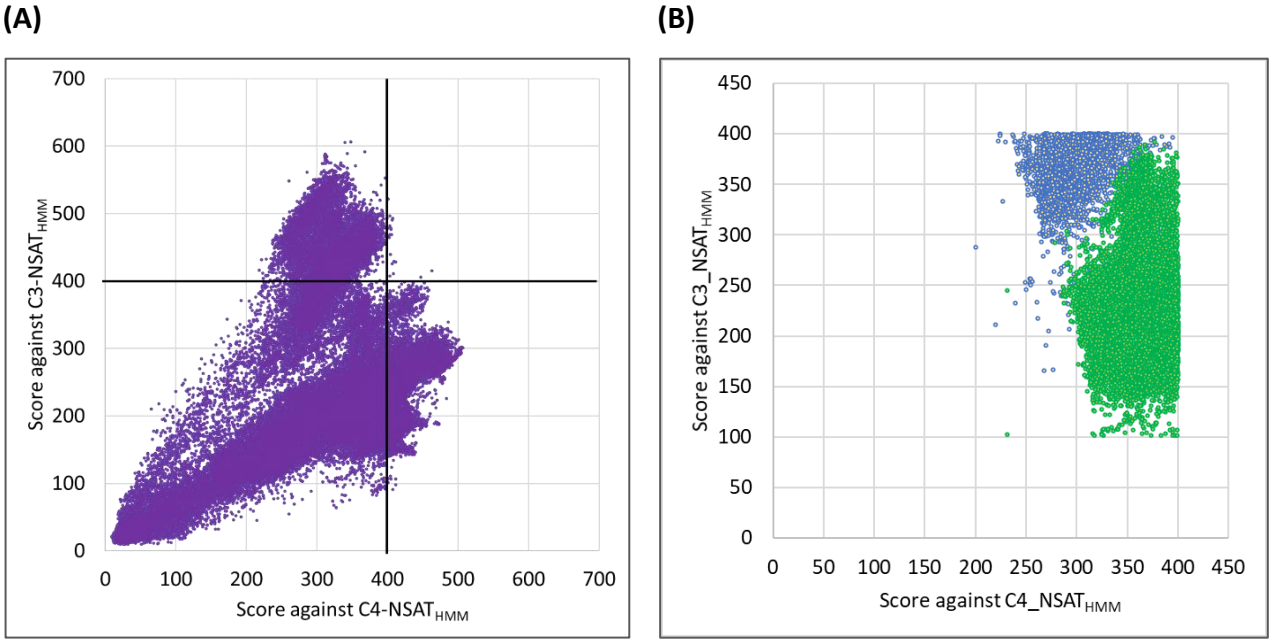

(A)
(B)
Score against C3-NSAT$_{HMM}$
Score against C4-NSAT$_{HMM}$
Score against C3_NSAT$_{HMM}$
Score against C4_NSAT$_{HMM}$

Figure 4

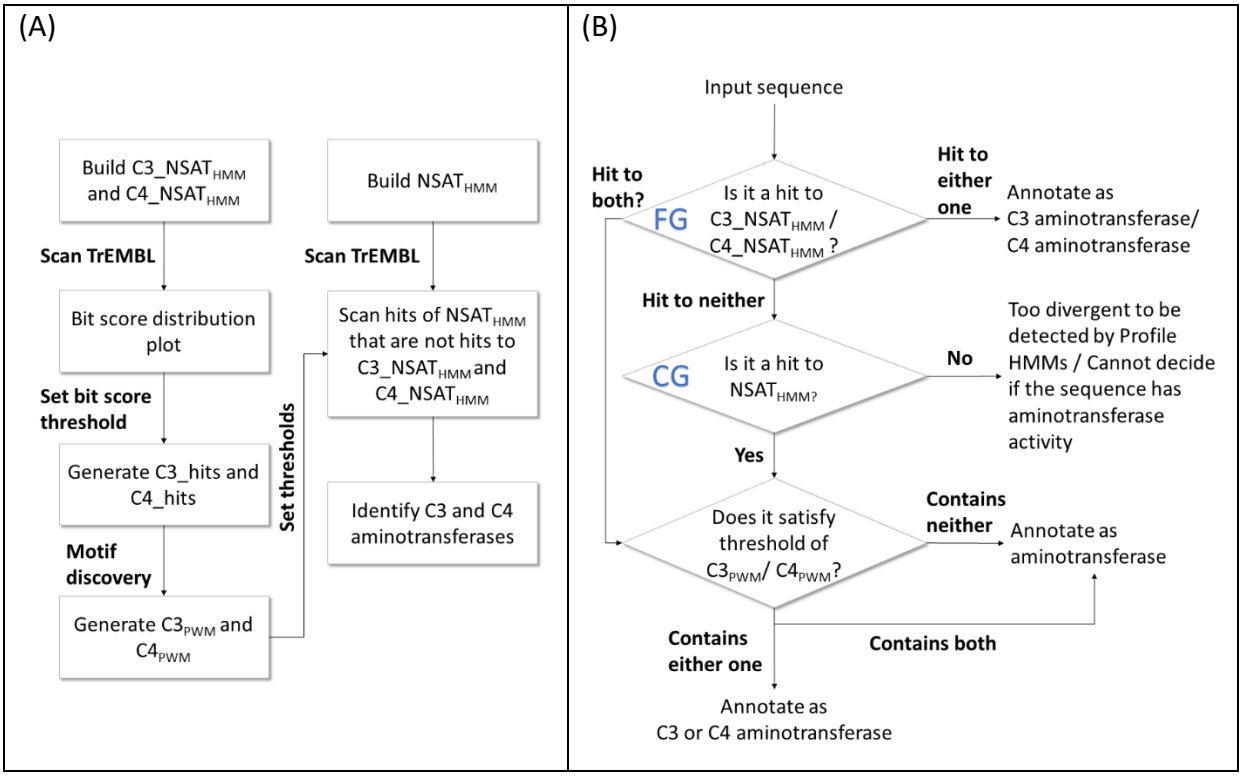

(A)
Build C3_NSAT$_{HMM}$ and C4_NSAT$_{HMM}$
Scan TrEMBL
Bit score distribution plot
Set bit score threshold
Generate C3_hits and C4_hits
Motif discovery
Generate C3$_{PWM}$ and C4$_{PWM}$
Set thresholds
Build NSAT$_{HMM}$
Scan TrEMBL
Scan hits of NSAT$_{HMM}$ that are not hits to C3_NSAT$_{HMM}$ and C4_NSAT$_{HMM}$
Identify C3 and C4 aminotransferases
(B)
Input sequence
FG
Is it a hit to C3_NSAT$_{HMM}$ / C4_NSAT$_{HMM}$?
Hit to both?
Hit to either one
Annotate as C3 aminotransferase/ C4 aminotransferase
Hit to neither
CG
Is it a hit to NSAT$_{HMM?}$
No
Too divergent to be detected by Profile HMMs / Cannot decide if the sequence has aminotransferase activity
Yes
Does it satisfy threshold of C3$_{PWM}$/ C4$_{PWM}$?
Contains neither
Annotate as aminotransferase
Contains either one
Contains both
Annotate as C3 or C4 aminotransferase

Figure 5

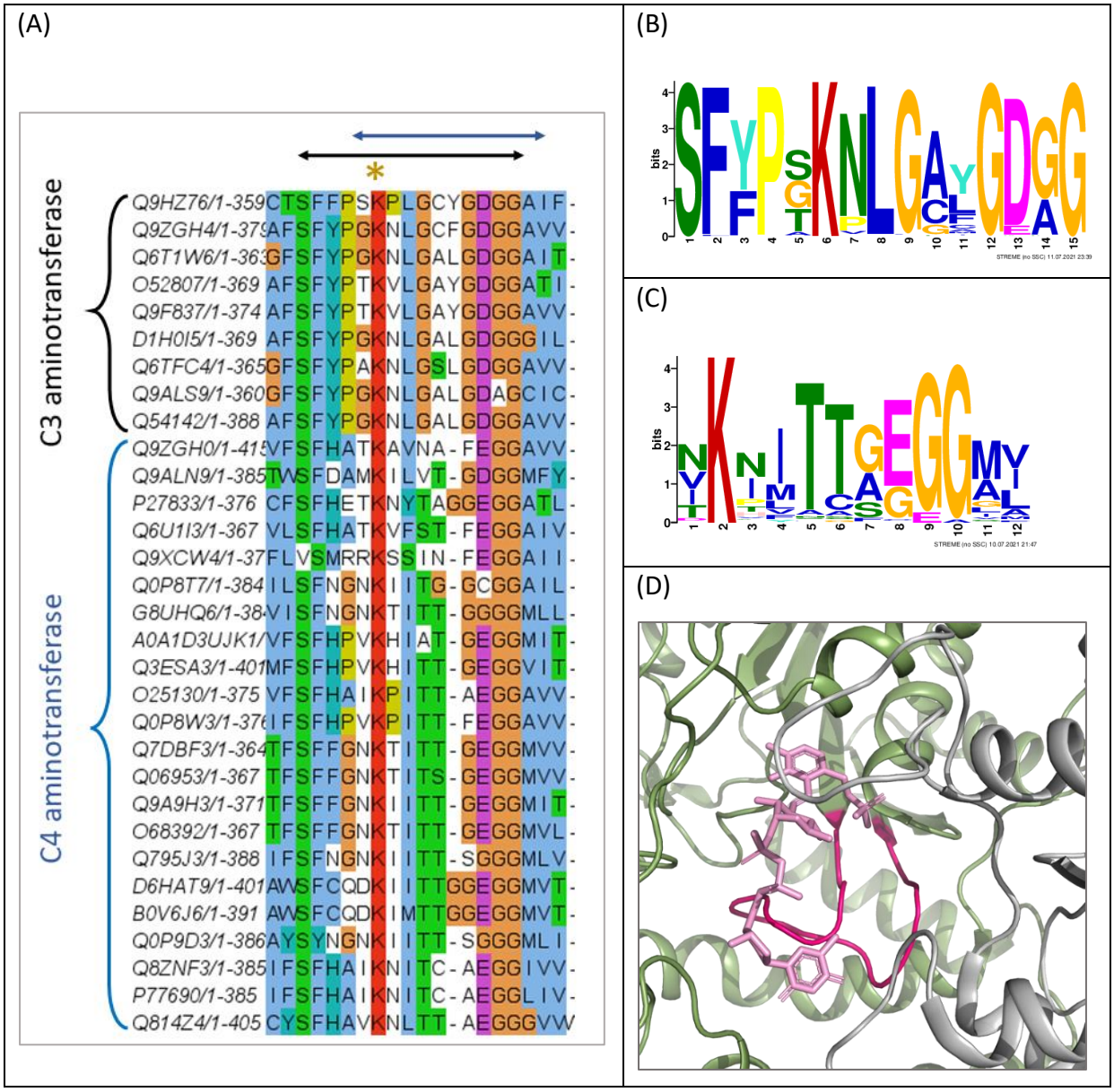

(A)
C3 aminotransferase
C4 aminotransferase
Q9HZ76/1-359 CTSFFPSKPLGCYGDGGAIF-
Q9ZGH4/1-379 AFSFYPGKNLGCFGDGGAVV-
Q6T1W6/1-363 GFSFYPGKNLGALGDGGAIT-
O52807/1-369 AFSFYPTKVLGAYGDGGATI-
Q9F837/1-374 AFSFYPTKVLGAYGDGGAVV-
D1H0I5/1-369 AFSFYPGKNLGALGDGGGIL-
Q6TFC4/1-365 GFSFYPAKNLGSLGDGGAVV-
Q9ALS9/1-360 GFSFYPGKNLGALGDAGCIC-
Q54142/1-388 AFSFYPGKNLGALGDGGAVV-
Q9ZGH0/1-415 VFSFHATKAVNA-FEGGAVV-
Q9ALN9/1-385 TWSFDAMKILVT-GDGGMFY-
P27833/1-376 CFSFHETKNYTAGGEGGATL-
Q6U1I3/1-367 VLSFHATKVFST-FEGGAIV-
Q9XCW4/1-37 FLVSMRRKSSIN-FEGGAII-
Q0P8T7/1-384 ILSFNGNKIITG-GCGGAIL-
G8UHQ6/1-38 VISFNGNKTITT-GGGGMLL-
A0A1D3UJK1/ VFSFHPVKHIAT-GEGGMIT-
Q3ESA3/1-401 MFSFHPVKHITT-GEGGVIT-
O25130/1-375 VFSFHAIKPITT-AEGGAVV-
Q0P8W3/1-376 IFSFHPVKPITT-FEGGAVV-
Q7DBF3/1-364 TFSFFGNKTITT-GEGGMVV-
Q06953/1-367 TFSFFGNKTITS-GEGGMVV-
Q9A9H3/1-371 TFSFFGNKIITT-GEGGMIT-
O68392/1-367 TFSFFGNKTITT-GEGGMVL-
Q795J3/1-388 IFSFNGNKIITT-SGGGMLV-
D6HAT9/1-401 AWSFCQDKIITTGGEGGMVT-
B0V6J6/1-391 AWSFCQDKIMTTGGEGGMVT-
Q0P9D3/1-386 AYSYNGNKIITT-SGGGMLI-
Q8ZNF3/1-385 IFSFHAIKNITC-AEGGIVV-
P77690/1-385 IFSFHAIKNITC-AEGGLIV-
Q814Z4/1-405 CYSFHAVKNLTT-AEGGGVW
(B)
bits
STREME (no SSC) 11.07.2021 23:39
(C)
bits
STREME (no SSC) 10.07.2021 21:47
(D)

Figure 6

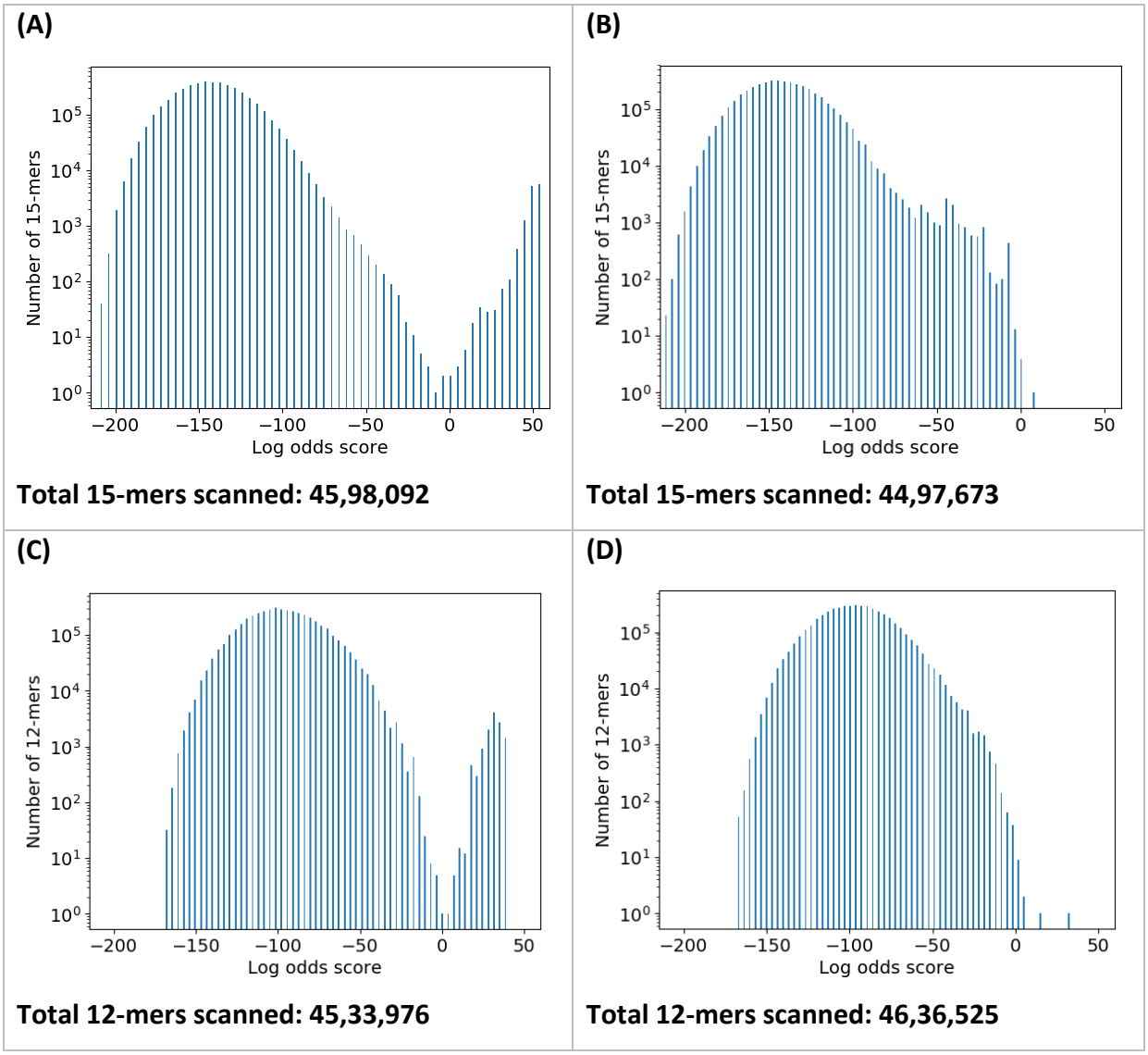

(A)
Number of 15-mers
Log odds score
Total 15-mers scanned: 45,98,092
(B)
Number of 15-mers
Log odds score
Total 15-mers scanned: 44,97,673
(C)
Number of 12-mers
Log odds score
Total 12-mers scanned: 45,33,976
(D)
Number of 12-mers
Log odds score
Total 12-mers scanned: 46,36,525

Figure 7