Searching sequence databases for functional homologs using profile HMMs: how to set bit score thresholds?

Jaya Srivastava*, Ritu Hembrom, Ankita Kumawat 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: HexNAc 6-dehydrogenase, hexose 6-dehydrogenase, alpha-lactalbumin, lysozyme, SH3 domain, GT-A fold

List of abbreviations

ARBA	Association-Rule-Based Annotator
BLAST	Basic Local Alignment Search Tool
GT	Glycosyltransferase
Hex-DH	NDP-hexose 6-dehydrogenases
HexNAc-DH	UDP-HexNAc 6-dehydrogenases
HMM	Hidden Markov Model
KO	KEGG Ortholog
MAFFT	Multiple Alignment using Fast Fourier Transform
MEME	Multiple Expectation maximizations for Motif Elicitation
MSA	Multiple Sequence Alignment
NDP 6-DH	Nucleotide sugar 6-dehydrogenase
NDP-Hex	NDP-hexose
PSI-BLAST	Position-specific iterative BLAST
ROC	Receiver Operating Characteristic
UDP-HexNAc	UDP-N-acetyl hexosamine

ABSTRACT

Motivation: UniProt and BFD databases together have 2.5 billion protein sequences. A large majority of these proteins have been electronically annotated. Automated annotation pipelines, vis-à-vis manual curation, have the advantage of scale and speed but are fraught with relatively higher error rates. This is because sequence homology does not necessarily translate to functional homology, molecular function specification is hierarchic and not all functional families have the same amount of experimental data that one can exploit for annotation. Consequently, customization of annotation workflow is inevitable to minimize annotation errors.

Results: We discuss possible ways of customizing the search of sequence databases for functional homologs using profile HMMs. Choosing an optimal bit score threshold is a critical step in the application of HMMs, which is illustrated using four Case Studies; the single domain nucleotide sugar 6-dehydrogenase and lysozyme-C families, and SH3 and GT-A domains which are typically found as a part of multi-domain proteins. We also discuss the limitations of using profile HMMs for functional annotation and suggests some possible ways to partially overcome such limitations.

Supplementary information

Supplementary_material containing Figures S1-S7 and Tables S1 and S2

Supplementary_dataset.xlsx

INTRODUCTION

Hidden Markov Model (HMM) is a statistical modelling approach that has gained success across diverse fields (Rabiner et al. 1983; Lander and Green 1987; Churchill 1989; Yang et al. 1994). Profile HMMs have been used for the identification of sequence homologs (Krogh, Brown, et al. 1994), prediction of coding regions (Krogh, Mian, et al. 1994), fold recognition (White et al. 1994; Hubbard and Park 1995), etc. Profile HMMs were introduced in sequence analysis as models to capture sequence variability in all homologous positions of a sequence family of proteins (Krogh, Brown, et al. 1994). Profile HMMs outperform pairwise BLAST in detecting remote homologs (Park et al. 1998) because BLAST does not have a mechanism to consider position specific sequence variations among homologs. Profile HMMs, like PSI-BLAST, use position-specific scores but in addition, they penalize gaps in conserved regions much more than those in variable regions (Durbin et al. 1998; Stojmirović et al. 2008). Consequently, profile HMM databases such as Pfam (Punta et al. 2012) and SUPERFAMILY (Gough et al. 2001) are being used for functional annotation of new protein sequences.

The effectiveness of an algorithm is critically dependent on the values one sets for its parameters. For algorithms designed to identify (remote) homologs, suggested default values are tuned to make the search sensitive, rather than specific. HMMER developers suggest E-value = 10 as the default threshold for *hmmsearch* (Finn et al. 2011) i.e., if one were to search a database of size S, even sequences which score t bits and E-value 10 will be output as hits. Clearly, this threshold leads to (very) low specificity and hence is unsuitable for sequence-similarity based functional annotation. Thus, a critical step in implementing profile HMMs for detecting functional homologs is the identification of an optimal E-value or bit-score threshold. Choosing an optimal threshold is not straightforward primarily because (i) sequence homology does not necessarily translate to functional homology and (ii) there are not adequate number of experimentally characterized sequences in every family of functional homologs. With an automated process for

setting bit score thresholds, optimality cannot be guaranteed since every family of functional homologs is unique. Instead, customizing a workflow taking into consideration the nature and amount of available experimental data can lead to higher prediction accuracies. Herein, we illustrate this approach using two functional families and two domains as Case Studies. It will not be out of place here to mention that we have a catch-22 situation on hand since limitations on the resources for performing experiments mandate the use of prediction methods and validation by experiments is the only way to know the veracity of (large scale) predictions.

METHODS

Databases and web servers used in this study are listed in Table S1. Input dataset for building profile HMMs consisted of experimentally characterized proteins and reviewed sequences from Swiss-Prot (Figure 1). This type of dataset is henceforth referred to as Exp_dataset. Exp_datasets for Case Studies 1 and 2 were obtained from literature (searched using Google Scholar) and Swiss-Prot, followed by manually verifying that activities have been tested by product characterization, direct enzyme assay, or gene disruption or complementation. The keywords GlcNAc / ManNAc / HexNAc / glucose / mannose / hexose/ sugar 6-dehydrogenases were used for Case Study 1, and Lysozyme C and alpha-lactalbumin for Case Study 2. Sequences that were missed out by keyword search were obtained by iterative search i.e., searching Swiss-Prot using an HMM built from an initial set of sequences.

Bit scores were used as thresholds rather than E-values since they remain the same irrespective of the size of the database searched. Exp_dataset sequences were scored against the corresponding profile and the lowest bit score for Exp_dataset sequences was set as T_{exp}. Entries in the TrEMBL database are automatically annotated using UniProtKB's UniRule and Association-Rule-Based Annotator (ARBA) (UniProt Consortium 2021). These entries were used to generate Receiver Operator Characteristic (ROC) curves, as described in more detail elsewhere (Srivastava et al. 2020). In Case Studies 1 and 2, which involved functional families,

hits to profiles were analysed for residue conservation using in-house python scripts. Such an analysis could not be done for Case Studies 3 and 4 since profiles in these Case Studies are meant to detect domains and domain boundaries rather than specific molecular functions.

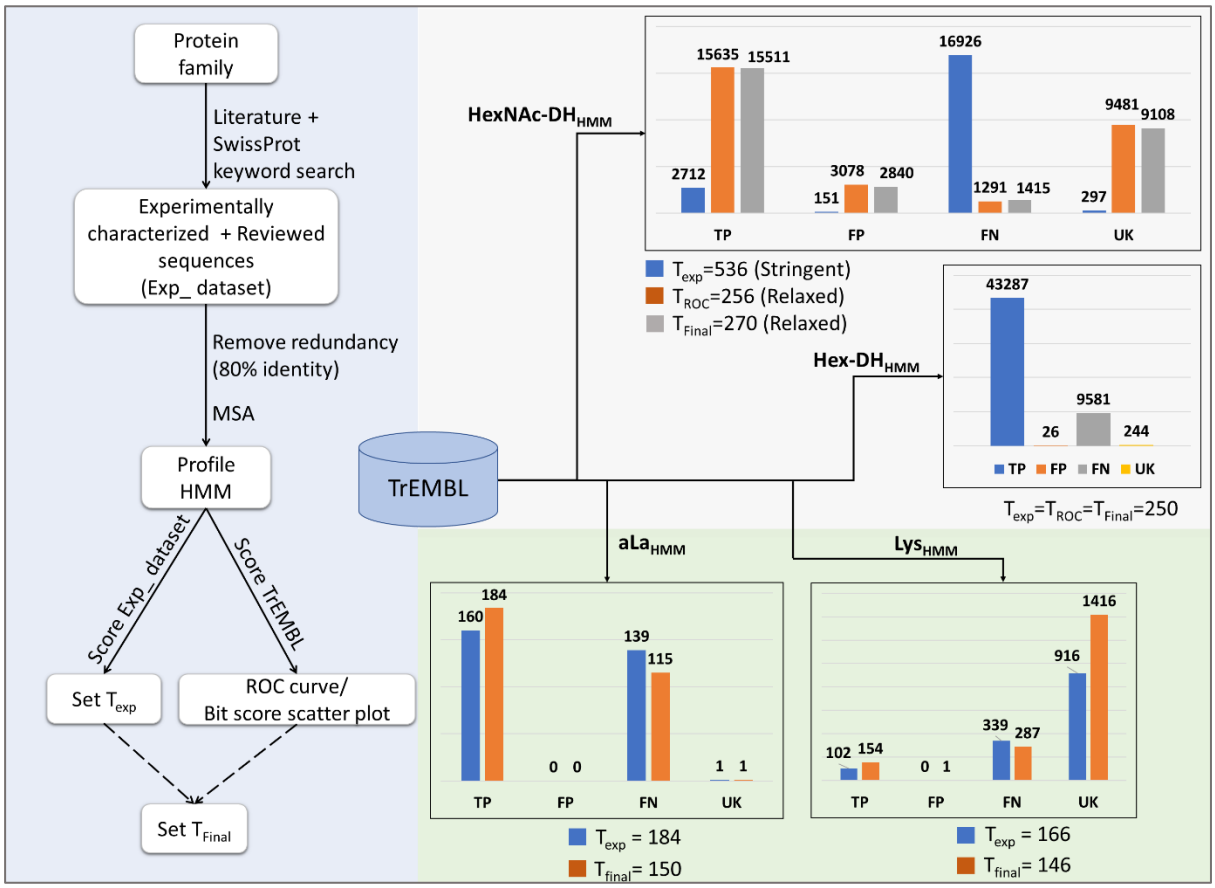


Figure 1: Procedure used to derive Profile HMMs and analyse hits for functional families in Case Studies 1 and 2. The final threshold, T_{Final}, can be the same as T_{exp} or T_{ROC}, or is set based on a bit score scatter plot. Details of Profile HMMs are given in Table 1. Details of proteins that constituted Exp_dataset are given in Supplementary_dataset.xlsx: NDP-6-DHs and lysozyme-C. TP and FN: True positives (bit score >= threshold) and false negatives (bit score < threshold) are proteins which have the same molecular function annotation as that assigned to the profile HMM. Here, threshold is T_{exp}, T_{ROC} or T_{Final}, as the case may be.

FP: False positives (bit score < threshold) are proteins which have a molecular function annotation that is different from the one that is assigned to the profile HMM.

UK: Proteins for which bit score >= threshold but there is incomplete or no molecular function annotation.

Table 1: Details of profile HMMs built for Case Studies 1 and 2

Functional family	Name of the HMM profile	R[&]	NR[&]	T_{exp}[¶]	Closest homologs[$]	Bit score of closest homolog	T_{final}*	Approach used to set threshold	Annotation assigned
UDP-HexNAc 6-dehydrogenase (HexNAc-DH)	HexNAc-DH_{HMM}	21	12	536	NDP-Hex 6-dehydrogenase	217	270	ROC + Bit score comparison	UDP-HexNAc 6-dehydrogenase
NDP-Hex 6-dehydrogenase (Hex-DH)	Hex-DH_{HMM}	53	20	250	UDP-HexNAc 6-dehydrogenase	236	250	Bit score comparison	NDP-Hex 6-dehydrogenase
Calcium binding lysozyme-C [Ca(+)-Lys] [‡]	Ca(+)-Lys_{HMM}	6	5	220	Non-calcium binding lysozyme-C[#]	199	Since T_{exp} for Ca(-)-Lys_{HMM} is less than the bit score for its closest homolog (viz., Ca(+)-Lys), these two datasets were merged together to form the dataset Lysozyme-C.		
Non-calcium-binding lysozyme-C [Ca(-)-Lys] [‡]	Ca(-)-Lys_{HMM}	87	22	169	Calcium binding lysozyme-C[#]	189			
Lysozyme-C	Lys_{HMM}	93	26	166	a-Lactalbumin	145	140	Bit score comparison	Lysozyme-C / Lysozyme-C-like
alpha-Lactalbumin	aLa_{HMM}	22	12	184	Lysozyme-C	140	150	Bit score comparison	alpha-Lactalbumin / alpha-Lactalbumin-like

[&] The number of sequences which have been experimentally characterized or reviewed in the Swiss-Prot database. The sequence identity cut-off was set to 80% to obtain the non-redundant **NR** dataset from the redundant dataset **R**. UniProt IDs and other descriptions are given in Supplementary_dataset.xlsx: NDP-6-DHs and lysozyme-C.

[¶] The lowest domain bit score for the dataset **R**.

[$] Closest homolog is the highest scoring false positive that appears as a hit to a profile HMM when the threshold is set to HMMER's default value.

* This is the final threshold recommended for the profile. Sequences satisfying this threshold are assigned annotations given in the column "Annotation assigned"

[#] Calcium binding property of lysozyme-C sequences was verified from literature.

[‡] Profile HMMs could not distinguish calcium-binding lysozymes from those which do not bind calcium: Both Ca(+)-Lys and Ca(-)-Lys retain catalytic activity. The ability to bind calcium is imparted by a 7-residue Asp-rich motif (Tsuge et al. 1992). Two profiles were generated by considering Ca(+)-Lys and Ca(-)-Lys separately. However, bit scores of both Ca(+)-Lys and Ca(-)-Lys against either of the two profiles are comparable, indicating that HMMs cannot distinguish Ca(+) and Ca(-)-Lys from each other based on available experimental data.

To search for a sequence motif in a protein, position-specific amino acid frequency matrices were computed from MSA of the motif and converted into a position weight matrix of log odds score viz., $\log_2 \frac{f_{obs}}{f_{exp}}$, where f_{obs} and f_{exp} are observed and expected frequencies of occurrence of amino acid residues in a column of the MSA (i.e., a position in a motif). Expected frequencies were fetched from the TrEMBL database (Release 2021_01). For a given sequence, the n-residue stretch (n is the length of the motif) that has the highest log odds score was taken to be the best possible match for the motif; indeed, highest scoring motifs were subsequently verified to be the motifs of interest using in-house python scripts.

RESULTS

Case Study 1: Nucleotide sugar 6-dehydrogenase families

Generating Profile HMMs: Nucleotide sugar 6-dehydrogenases include HexNAc-DH and Hex-DH (Figure 2). They belong to the Short-chain Dehydrogenase Reductase (SDR) superfamily (Borg et al. 2020) and catalyze the conversion of hexose / HexNAc to the corresponding uronic acid. Two profiles, HexNAc-DH_{HMM} and Hex-DH_{HMM}, were built to identify proteins belonging to these two functional families (Figure 1).

ROC curves: The TrEMBL database has nearly 17,000 sequences that are annotated as one of the HexNAc-DHs. Scoring TrEMBL using HexNAc-DH_{HMM} with T_{exp} (Table 1) as the threshold suggested that T_{exp} has poor sensitivity (Figure 1). To improve sensitivity, threshold was set to 256 bits (T_{ROC}) using an ROC curve (Figure 3). With this threshold, there was a marked increase in sensitivity but with a concomitant increase in false positives (Figure 1). TrEMBL was scored using Hex-DH_{HMM} also with T_{exp} (Table 1) as the threshold. 81% of ~53000 sequences annotated as one of the Hex-DHs in TrEMBL were obtained as hits along with a very small number of sequences as false positives; this implied that T_{exp} is optimal for Hex-DH_{HMM} (Figure 1). An ROC curve was generated to find a threshold with a higher sensitivity but for thresholds less than T_{exp}, specificity decreased without much improvement in sensitivity (Figure 3).

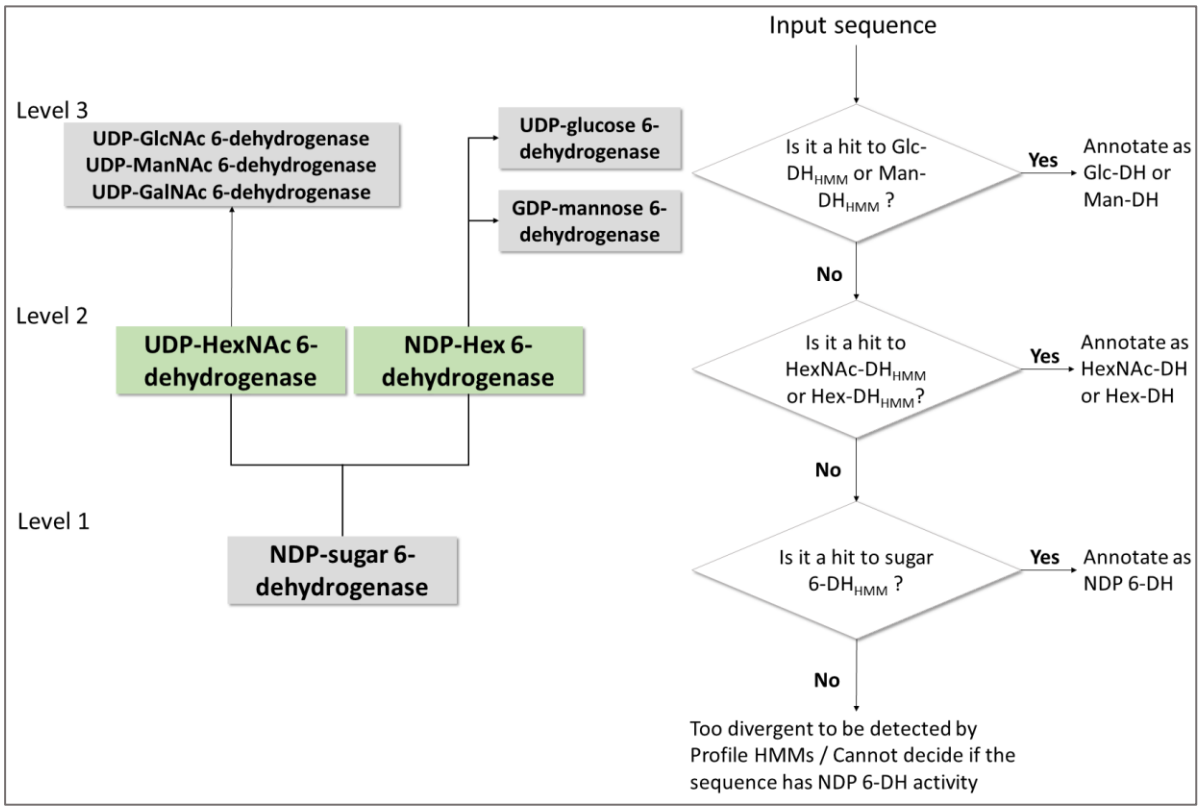


Figure 2: Categorization of nucleotide sugar 6-dehydrogenase family proteins into different levels based on substrate specificity. Assignment of substrate specificity is generic in Level 1 and increasingly specific in Levels 2 and 3. Separate profile HMMs could not be generated for the three HexNAc 6-DHs because of paucity of data and hence only a combined profile viz., HexNAc-DH$_{HMM}$ could be generated. Rules on the right illustrate a stepwise approach to annotate an input sequence using profile HMMs built at three levels of categorization. Families whose profile HMMs are discussed in this study are shown in green boxes.

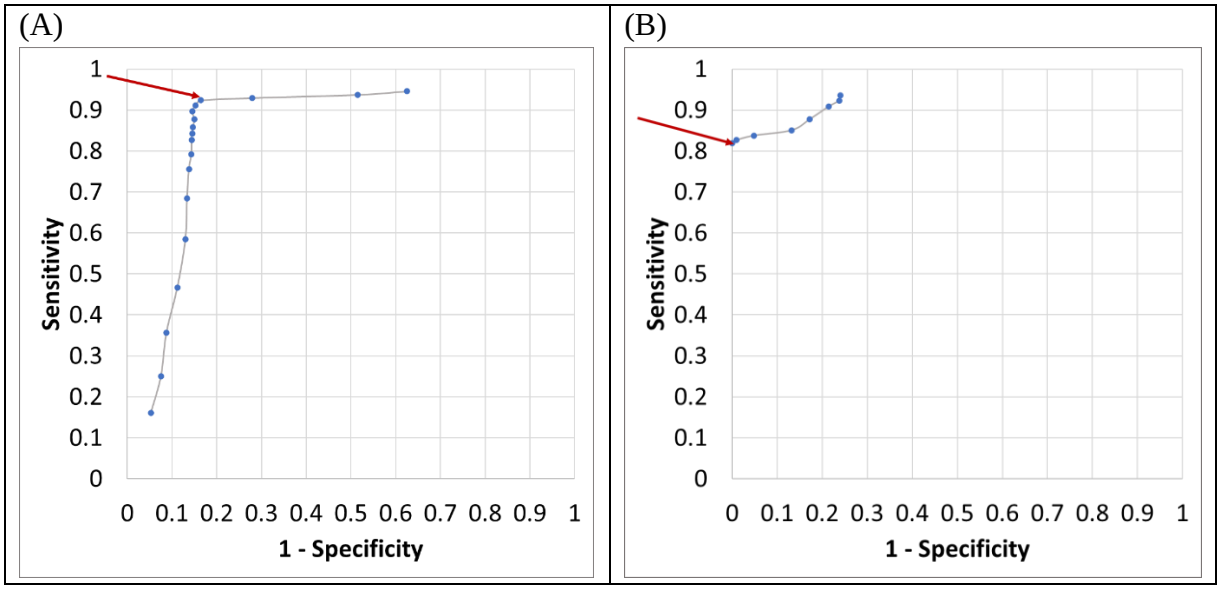


Figure 3: ROC curves for (A) HexNAc-DH_{HMM} and (B) Hex-DH_{HMM} obtained by scoring the TrEMBL database. Data points have been merely joined by a grey line. Arrow points to the data point correspond to the bit score threshold determined from the curve viz., T_{ROC}.

Discriminating HexNAc-DH from Hex-DH: HexNAc-DH and Hex-DH are the closest homologs of each other (Table 1). 2800 out of 2840 false positives for the HexNAc-DH_{HMM} from TrEMBL are annotated as one of the Hex-DHs. To find-tune the threshold to improve specificity, TrEMBL was scored using both HexNAc-DH_{HMM} and Hex-DH_{HMM} using default values suggested by HMMER as thresholds. Plotting bit scores of hits common to the two profiles (Figure 4) showed that proteins which are "high-scoring" for one profile are "low-scoring" for the other. This plot showed that T_{exp} (250 bits) itself is optimal for Hex-DH_{HMM}; however, T_{Final} for HexNAc-DH_{HMM} was revised marginally viz., from 256 to 270 bits (Table 1). Expectedly, this revision resulted in a marginal increase in specificity and decrease in sensitivity. Barring a few, false positives of HexNAc-DH_{HMM} score higher against HexNAc-DH_{HMM} as against Hex-DH_{HMM}. Additionally, 98% of these false positives retain residues shown to be essential for catalysis in the HexNAc-DH family. Are these proteins actually HexNAc-DHs but are incorrectly annotated as Hex-DHs in TrEMBL?

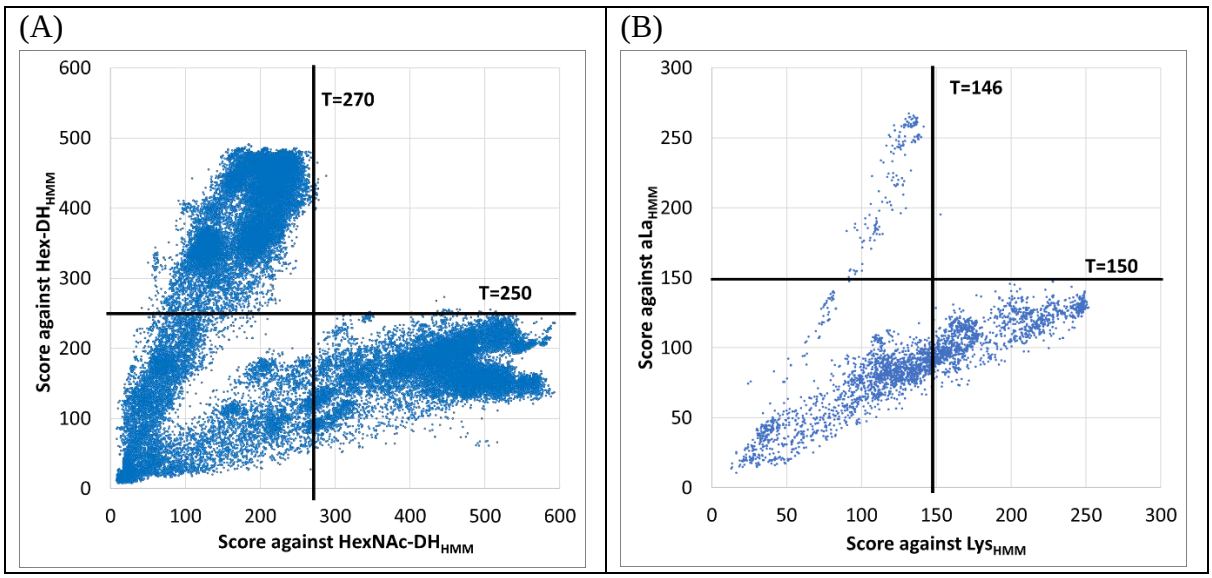


Figure 4: (A) Bit score comparison of TrEMBL entries that are hits to both Hex-DH$_{HMM}$ and HexNAc-DH$_{HMM}$. (B) Bit score comparison of common hits from TrEMBL against Lys$_{HMM}$ and aLa$_{HMM}$.

Residue conservation analysis: Conservation of active site residues among hits was examined to analyze the optimality of the threshold. In HexNAc-DHs, the cofactor binding motif GxGxxG, conserved cysteine, arginine and lysine (Lys204, Arg211 and Cys258, *Staphylococcus aureus* UDP-ManNAc 6-dehydrogenase numbering (Gruszczyk et al. 2011)) are implicated as important for catalysis. These are conserved in ~96% of the sequences. Similarly, active site residues of NDP-hexose 6-dehydrogenases (Glu161, Lys210, Asn214, Lys218, Cys268 and Asp272, *Pseudomonas aeruginosa* GDP-mannose 6-dehydrogenase numbering) (Snook et al. 2003; Rocha et al. 2011) are conserved in ~95% of the hits. Bit scores of sequences in which these residues are absent span the entire range. For instance, the second glycine of the GxGxxG motif is absent in 601 hits of HexNAc-DH$_{HMM}$, whose bit scores range from 270-539. These observations indicate that profile HMMs are insensitive to the absence of critical residues i.e., the bit score can be higher than the threshold even when the mutation happens to be in a conserved region. Nevertheless, the number of false positives based on residue conservation are very low, suggesting that the threshold derived from aforementioned approaches is reliable. Thus,

combining profile HMM with bit score distribution plot helped in discriminating families with different substrate specificities.

Fine-grained function annotation: HexNAc-DHs and Hex-DHs belong to the SDR superfamily and share the same fold (Gruszczyk et al. 2011; Rocha et al. 2011); none of the other functional families appear as hits to these profiles. Profile HMMs are able to capture sequence variations that give rise to utilization of either UDP-HexNAc or NDP-Hex. Using a similar approach as described above, we were also able to differentiate between UDP-glucose 6-dehydrogenases and GDP-mannose 6-dehydrogenases, enzymes that are strictly specific towards the respective substrates (Srivastava et al. 2020). However, it was challenging to resolve the substrate specificity among HexNAc-DHs because (i) some of these enzymes can utilize more than one substrate (Zhao et al. 2000) and (ii) experimental data on substrate specificity is available for very few HexNAc-DHs. Therefore, homologs of these enzymes identified using profile HMMs could, at best, be annotated as UDP-HexNAc 6-dehydrogenases. In scenarios where substrate specificity cannot be resolved using these profiles, an additional profile for all NDP 6-DHs can be used to capture homologs that are not hits to substrate-specific profiles. Using this set of profiles, rules for annotation can be formulated for an input sequence by scoring it at different levels, starting from the profile with highest substrate specificity, i.e., level 3 (Figure 2).

Case Study 2: Lysozyme-C family

Sequence and structure homologs with mutually exclusive functions: Lysozyme-C family includes lysozymes, lysozyme-like proteins and alpha-lactalbumins (Prager and Wilson 1988; Grobler, Rao, et al. 1994; Irwin et al. 2011; Kalra et al. 2016). Some lysozymes bind calcium whereas others do not, denoted henceforth as Ca(+)-Lys and Ca(-)-Lys, respectively. alpha-Lactalbumins have lost catalytic activity but have acquired a new function wherein, as part of the lactose synthase complex, they modify the acceptor substrate specificity of beta-1,4-galactosyltransferase (Ramakrishnan and Qasba 2001). Phylogenetic analyses have indicated that

alpha-lactalbumins and lysozymes diverged from a common ancestor, a calcium binding protein (Prager and Wilson 1988; Irwin et al. 2011) (Figure S1).

Lysozyme-C family members exhibit high sequence and 3D structure similarities (Qasba and Kumar 1997) (Figure S2) despite of functional divergence. alpha-Lactalbumins (Prager and Wilson 1988) and lysozyme-like proteins (Irwin et al. 2011) share at least 30-40% sequence identity with lysozymes; this precludes the use of sequence identity alone for function annotation. Herein, we present a procedure to set bit score thresholds to distinguish alpha-lactalbumins from other members of this family.

Setting thresholds: Two profiles viz., Lys_{HMM} and aLa_{HMM} were generated for lysozyme and alpha-lactalbumins (Figure 1; Table 1). As mentioned earlier, T_{exp} led to poor sensitivity in the case of HexNAc-DH_{HMM}, which meant that the threshold had to be lowered significantly. Is there a need for downward revision of T_{exp} of Lys_{HMM} and aLa_{HMM} also? It is noted that the highest score of an alpha-lactalbumin against Lys_{HMM} is 145 bits; similarly, the highest score of a lysozyme against aLa_{HMM} is 140 bits (Table 1). An ROC curve could not be plotted for Lys_{HMM} as there were no false positives scoring >145 bits among its hits from TrEMBL. Therefore, we used bit score comparison approach to set thresholds for both Lys_{HMM} and aLa_{HMM} (Figure 4). Accordingly, T_{Final} was set to 146 bits for Lys_{HMM} and 150 bits for aLa_{HMM} (Table 1). Case Studies 1 and 2 collectively show that a set of test sequences are required to adjust bit score threshold for profile HMMs that are built to annotate functions at fine grained level (Figure 2, Figure S1B).

Analysis of annotations of TrEMBL hits to Lys_{HMM}: Using T_{Final} (Table 1) as the threshold, 1571 hits were obtained from TrEMBL. Of these, very few are annotated as lysozyme-C; many have incomplete or ambiguous annotations ("lysozyme", "alpha-lactalbumin/lysozyme", "lysozyme-like") or are annotated as 'sperm acrosome membrane-associated protein', or do not have

molecular function annotation (Figure 1). Several of these hits do not have the catalytic residues Glu53 and Asp70 (chicken egg-white lysozyme numbering (Malcolm et al. 1989)). Scores of such sequences fall in the range 146-236 bits. Sperm acrosome membrane-associate proteins are described in literature as lysozyme-like proteins, some of which have lost their bacteriolytic activity due to catalytic site mutations (Herrero et al. 2005; Wei et al. 2013: 6; Zhang et al. 2016). As observed earlier, profile HMMs are insensitive to point mutations and thus, catalytically active proteins could not be distinguished from inactive homologs using HMMs. To accommodate this limitation, the annotation 'lysozyme-C / lysozyme-C-like' was assigned for Lys_{HMM} (Table 1).

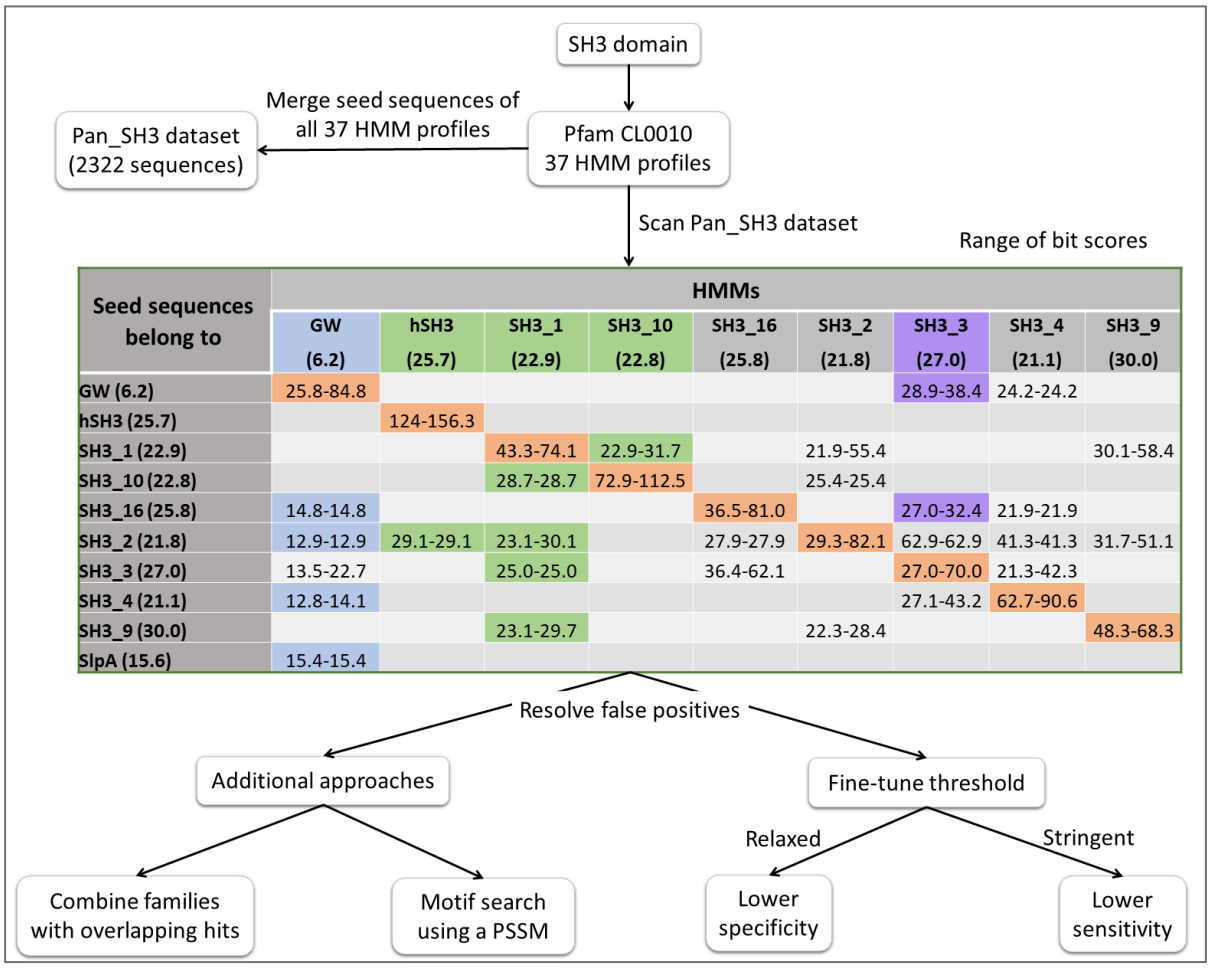
Analysis of annotations of TrEMBL hits to aLa_{HMM}: Using T_{Final} (Table 1) as the threshold, 185 TrEMBL entries are obtained as hits, of which one has no molecular function annotation (Figure 1). 105 sequences are annotated as 'Lactose synthase B' and hence are true positives. Sequences of true positives were analysed for conservation of functionally important residues viz., Phe50, His51, Gln138 and Trp137 (bovine alpha-lactalbumin numbering (Grobler, Wang, et al. 1994)). Phe50 and His51 are conserved among all hits but Gln136 and Trp137, residues implicated in binding to galactosyltransferase (Grobler, Wang, et al. 1994), are absent in 55 sequences due to ~20-30 residue C-terminal truncation. Such sequences score in the range 150-222 bits. The effect of such a truncation on lactose synthase activity is not known. Hence, the annotation 'alpha-lactalbumin / alpha-actalbumin-like' was assigned for the aLa_{HMM} (Table 1).

Case Study 3: SH3 domain

A highly divergent sequence family: Src homology 3 (SH3) domains are ~60-residue long structural elements characterized by a conserved beta-barrel fold (Figure S3) but no conserved sequence motif is found across all SH3 domains. Consequently, the Pfam database has 37 distinct families representing SH3 domains (Punta et al. 2012) (Supplementary_dataset.xlsx: SH3

families). SH3 domains are a part of multi-domain proteins that mediate diverse biological processes (Kurochkina and Guha 2013). Consequently, in databases such as TrEMBL, annotations for SH3 domain-containing proteins do not necessarily reflect the presence of this domain. Hence, unlike as in the case of single domain proteins, annotation-based approaches such as ROC curve are not suitable for defining bit score thresholds for SH3 domain profiles. Pfam assigns manually curated and periodically updated full-sequence and domain thresholds for every profile HMM, referred to as gathering thresholds by Pfam (Punta et al. 2012). Herein, we examine the domain gathering thresholds of SH3 domain profile HMMs (as of May 2021).

Are profile HMMs for all 37 families specific? A pan-SH3 sequence dataset was created by pooling the seed alignment sequences of all 37 profiles. Scoring this dataset against all 37 profiles showed that some of the sequences satisfy the gathering threshold of more than one profile (Figure 5) implying that their gathering thresholds are less specific. Such overlaps, especially among families within a clan, are proposed to reflect evolutionary relationships between families (Punta et al. 2012). Nonetheless, they can be resolved by further optimization of thresholds or identification of additional approaches that can be used to distinguish between closely related families. Pfam does revise gathering threshold periodically (Punta et al. 2012).



Seed sequences belong to	HMMs								
	GW (6.2)	hSH3 (25.7)	SH3_1 (22.9)	SH3_10 (22.8)	SH3_16 (25.8)	SH3_2 (21.8)	SH3_3 (27.0)	SH3_4 (21.1)	SH3_9 (30.0)
GW (6.2)	25.8-84.8						28.9-38.4	24.2-24.2	
hSH3 (25.7)		124-156.3							
SH3_1 (22.9)			43.3-74.1	22.9-31.7		21.9-55.4			30.1-58.4
SH3_10 (22.8)			28.7-28.7	72.9-112.5		25.4-25.4			
SH3_16 (25.8)	14.8-14.8				36.5-81.0		27.0-32.4	21.9-21.9	
SH3_2 (21.8)	12.9-12.9	29.1-29.1	23.1-30.1		27.9-27.9	29.3-82.1	62.9-62.9	41.3-41.3	31.7-51.1
SH3_3 (27.0)	13.5-22.7		25.0-25.0		36.4-62.1		27.0-70.0	21.3-42.3	
SH3_4 (21.1)	12.8-14.1						27.1-43.2	62.7-90.6	
SH3_9 (30.0)			23.1-29.7			22.3-28.4			48.3-68.3
SlpA (15.6)	15.4-15.4								

Figure 5: Steps used to analyse thresholds prescribed by Pfam for the 37 Profile HMMs for the SH3 domain. Seed sequences of all 37 profiles were scored against all 37 profiles. One or more seed sequences of 10 profiles (listed as rows) score above the domain gathering threshold of one or more profile (listed as columns) to which they do not belong (viz., false positives). Range of scores of seed sequences of a profile against own profile is highlighted in orange. Possible ways to resolve false positives are also depicted. Domain gathering thresholds of profiles are given in parenthesis following the name of the profile. Revising the gathering thresholds of hSH3 (from 25.7 bits to 30.0), SH3_1 (from 22.9 to 31) and SH3_10 (from 22.8 to 32 bits) excludes false positives (highlighted in green). For the profile GW (highlighted in blue), choosing the full sequence gathering threshold (21 bits) instead of the domain gathering threshold (6.2 bits) suffices to exclude most of the false positives. However, in the case of SH3_3 (highlighted in purple), a similar upward revision (from 27 to 38) leads to lower sensitivity concomitant with higher specificity. This holds true for SH3_16, SH3_2, SH3_4 and SH3_9 as well. The sequence Q5KV47_GEOKA/411-472 from the GW family scores 25.8 bits against GW profile but 38.4 and 24.2 bits against SH3_3 and SH3_4 profiles, respectively. The sequence coverage is higher against SH3_3 and SH3_4 profiles whereas it has long gaps in the seed alignment of GW family raising the possibility that is belongs to SH3_3 or SH3_4.

Revising thresholds of families with overlapping hits: In the case of the Gly-Trp (GW) domain profile, using the gathering threshold for the full sequence instead of that for the domain (https://pfam.xfam.org/family/PF13457#tabview=tab6) prevents sequence of SH3_2, SH3_4 and SH3_16 families from appearing as false positives to the GW profile (Figure 5). A similar upward revision of thresholds reduces false positives for SH3_1, hSH3 and SH3_10 also. However, such a revision compromises with the sensitivity of the profile in some cases (Figure 5). Under such a scenario, one may consider using two thresholds: (i) gathering threshold used by Pfam that has lower specificity (i.e., seed sequences of other SH3 families are hits) and assign a generic annotation viz., "SH3_1/hSH3/SH3_10/Gly-Trp domain" or (ii) use a revised, more specific threshold. For the remaining five families, multiple hits with scores comparable to that of seed sequences were obtained (Figure 5). To resolve false positives among hits of their profile HMMs, one can either employ residue/motif conservation analyses, or merge the seed sequences into a single profile and revise the annotation assigned to the profile (Figure 5). We will describe the motif conservation approach below.

Using family specific motifs to discriminate SH3 families: For this purpose, SH3_1 and SH3_2 families were considered as an illustrative example. Nearly 80% of SH3_1 seed sequences are hits to SH3_2 profile (Figure S4). Visual inspection of MSAs of seed sequences revealed a 6-residue extension in SH3_2 (Figure S5). To identify distinguishing motifs, SH3_1 and SH3_2 seed sequences were collectively submitted as input to MEME suite. Three motifs from the SH3_1 family are conserved in all the seed sequences, but their log-odds scores are inadequate to distinguish SH3_2 from SH3_1, either individually or in combination. Two motifs were identified in the SH3_2 family and these are conserved in 27 out of 28 seed sequences. One of these motifs is able to distinguish SH3_1 from SH3_2, with the exception of a false negative (the 28th sequence in which motif is absent). Log-odds scores of seed sequences of both families obtained from the aforementioned five motifs are given as supplementary data

(supplementary_dataset.xlsx: Log odds scores). Thus, by combining profile and motif search approach we can attain distinction between families that cannot be resolved based on profile HMMs alone.

Case Study 4: GT-A fold of glycosyltransferases

GT-A fold: Glycosyltransferases (GTs) are enzymes that catalyse the transfer of sugar moieties from activated donor molecules to acceptor molecules forming glycosidic bonds (http://www.cazy.org/GlycosylTransferases.html). CAZy has nearly 0.83 million GTs which are grouped into 114 sequence-based families (as of May 2021) and GT2 is the largest of these families with ~250,000 proteins. Most of the GTs whose 3D structure has been determined experimentally adopt either the GT-A or GT-B fold; few others have GT-C, GT-D or GT-E fold (Gloster 2014; Zhang et al. 2014; Kattke et al. 2019). CAZy has annotated that 19 GT families adopt GT-A fold including GT2. This is a similar scenario to that of SH3 family wherein 37 subfamilies are formed by Pfam based on sequence similarity, all of which adopt the SH3 fold.

Function annotation of multi-domain proteins: 3D structures of several GT2 family members have been determined (Table S2). It can be seen from these studies that the lower limit for the number of amino acids in a domain that adopts the GT-A fold (henceforth referred to as the GT-A domain) is around 210 residues. There are ~500 sequences in SwissProt that are labelled as GT2. The lengths of these proteins vary over a large range (Figure S6); this clearly suggests that many of them have one or more domains in addition to the GT-A domain. Functional annotation of such proteins necessarily involves the demarcation of domain boundaries, functional annotation of each domain as well as annotating the function performed by these covalently linked domains in unison. As the first step in this direction, since CAZy has assigned GT-A fold to GT2 family, a profile HMM-based approach was taken up to identify (i.e., find and demarcate domain boundary) GT-A domains in the GT2 family proteins (Figure 6).

MSA of GT-A fold domains: As mentioned earlier, 3D structures of 11 GT2 family proteins are known (Table S1). They vary in length from 256 to 978 amino acids, and thus, some are multi-domain proteins. GT-A domain boundaries in these proteins were identified from literature. Of these 11 proteins, chondroitin polymerase from *Escherichia coli* strain K4 (PDB id 2Z87) has two occurrences of the GT-A fold and these were taken as two separate GT-A domains. In all, the 12 GT-A domains were superimposed on each other using the Dali server (Figure 6). An MSA was derived from superimposed structures and this was used to generate GT-A_{HMM} (Figure 6). As has already been reported, MSAs derived from 3D structure superimposition are better than those derived solely from amino acid sequences (Carpentier and Chomilier 2019).

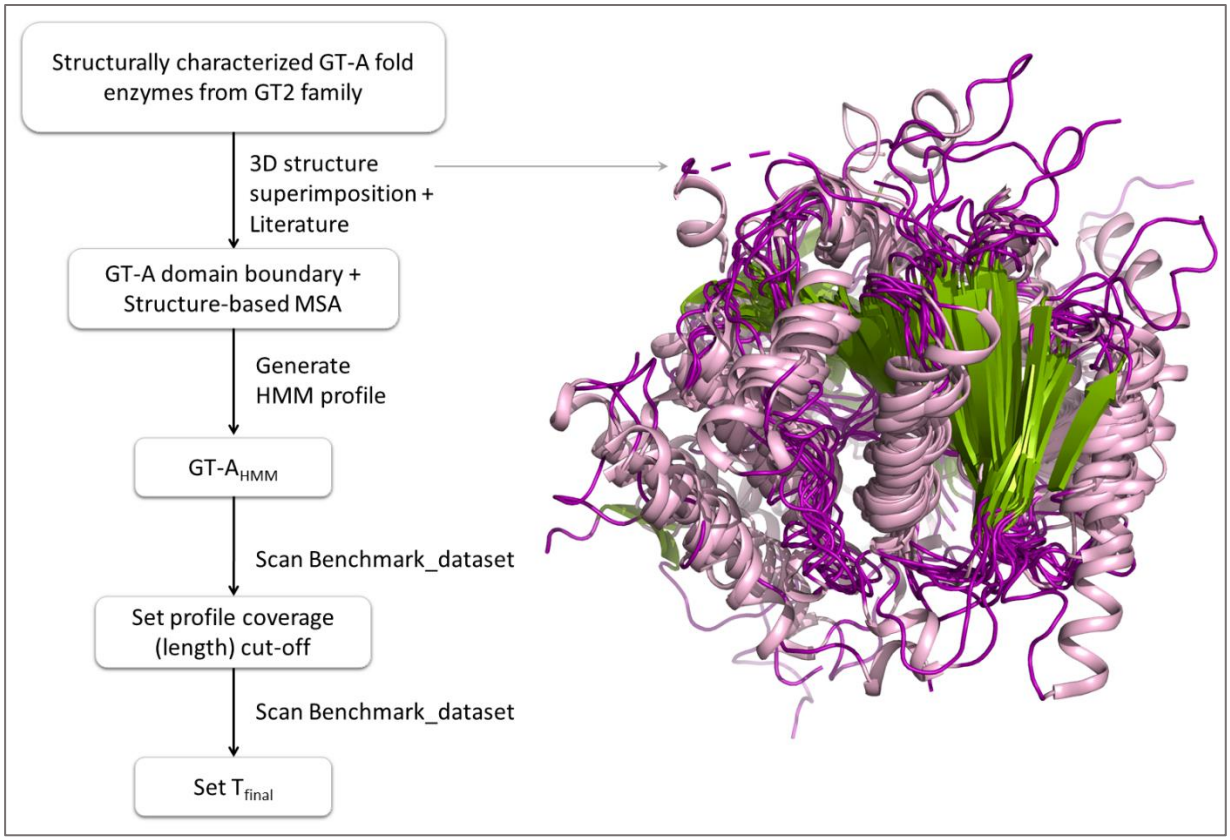


Figure 6: Steps followed to determine bit score threshold and HMM coverage cut-off for GT-A_{HMM}. PDB Ids and related information for proteins used for superposition are given in Table S1, Supplementary_dataset.xlsx: GT2 structures.

How to set bit score threshold for the GT-A_{HMM}? As in the case of SH3 domain containing proteins, GT-A domain containing proteins in the TrEMBL database are not expected to contain fold level annotations and hence using an ROC curve did not seem feasible. A curated dataset of sequences with experimental evidence (viz., enzyme activity assay, complementation studies, etc.) or those that have high sequence similarity to experimentally characterized sequences were chosen out of entries labelled as GT2 in SwissProt. This resulted in a benchmark dataset of 170 proteins (Supplementary_dataset.xlsx: GT2 benchmark dataset). The benchmark dataset was scored using GT-A_{HMM} with default threshold of HMMER. 137 sequences scored, whose domain bit score is in the range 3-172 with HMM coverage being as low as 15% (Supplementary_ dataset.xlsx: GT-A_{HMM} hits). However, bit score does not correlate with HMM coverage (Figure 7). HMM coverage of some of these sequences is <50 %, and the 33 sequences of the benchmark dataset do not score at all. Thus, it remains uncertain whether they adopt a GT-A fold. On the other hand, several hits have >70% coverage but score low, implying that they have diverged significantly while keeping the fold intact. These sequences do not match the C-terminal region of profile HMM (Supplementary_dataset.xlsx: GT-A_{HMM} hits). It is observed that from the MSA (Figure S7) that the C-terminal region (~20 residues) of sequences used to build GT-A_{HMM} are less conserved. We therefore set 160 residues as the HMM coverage cut-off (~78% of the length of the HMM profile). The lowest bit score of a sequence satisfying coverage cut-off is 25 bits, which was set as T_{Final} (Figure 7).

GT-A fold domain in other GT families of CAZy: GT-A_{HMM} was used to find domain homologs in other CAZy families that are annotated by CAZy as containing GT-A fold (GT6, GT7, GT8, GT15, GT27, GT64, GT81, etc.). Structurally characterized members of these GT families were scored using T_{Final}. We found a sequence (UniProt accession number Q86SR1) belonging to GT27 satisfies both coverage cut-off and T_{Final} (Supplementary_dataset.xlsx: GT-A_{HMM} hits), suggesting that the GT-A domain of this 603 residue protein shares sequence similarity with the

GT-A domain found in several GT2 family sequences. No other hits were obtained, implying that multiple profile HMMs are required to capture GT-A folds, as is the case with SH3 domains.

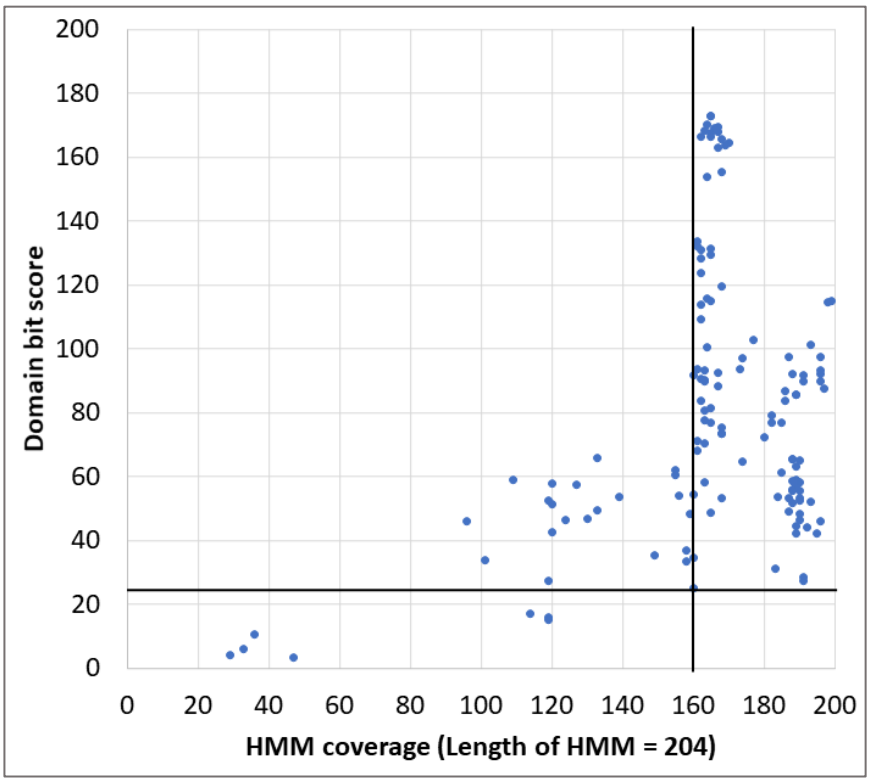


Figure 7: Bit scores of 133 sequences from GT2 benchmark dataset plotted against the corresponding HMM coverage (Supplementary_dataset.xlsx: GT-A_{HMM} hits). The horizontal line represents bit score cut off (=25 bits) and the vertical line represents HMM coverage cut-off (=160).

Discussion

Profile HMMs are powerful tools to detect remote homologs in sequence databases and sequenced genomes. The default threshold set by HMMER search tool is an E-value of 10, which most often results in poor specificity. The false positives can be similar sequences that have diverged to perform different functions, or homologous fragments present as part of other polypeptides. Therefore, using profile HMMs for the identification of homologs necessitates manual intervention to set thresholds that can increase specificity. We have used four Case Studies to illustrate approaches that can be adopted to optimize bit score thresholds.

Case Studies 1 and 2 describe approaches to identify homologs of functional families. We set a threshold for the nucleotide sugar 6-dehydrogenase family using a combination of ROC curve and bit score comparison. The ROC curve approach relies on molecular function annotations of TrEMBL sequences. Assuming that the annotations are correct, ~10% of the hits were found to be false positives against HexNAc-DH_{HMM} and almost all of these are annotated as Hex-DH. Almost all the false positives of HexNAc-DH_{HMM} scored more against HexNAc-DH_{HMM} than against Hex-DH_{HMM}, and retain active site residues of HexNAc-DH. On the other hand, only a small fraction of hits (i.e., 26 out of 43,557 hits (Figure 1)) appeared as false positives against Hex-DH_{HMM}. In the lysozyme-C family, we did not find any false positives against Lys_{HMM} and aLa_{HMM}, but we did observe several hits of Lys_{HMM} missing or having incomplete annotations (Figure 1). Such a skew in false positive rate is a consequence of the bias in the TrEMBL database from the view point of the extent of representation of different functional families. Another reason could be that even though TrEMBL annotations are largely correct, some sequences are incorrectly annotated and this warrants curation of annotation using additional approaches such as bit score comparisons. Such a curation can also lead to assigning specific annotations to entries with incomplete or missing annotations as is exemplified in the case of HexNAc-DH family (Figure 1). We also learn that in situations where substrate specificity cannot be assigned due to paucity of data, e.g., ManNAc 6-dehydrogenases or GlcNAc 6-dehydrogenases, a combined profile with broad annotation, i.e., HexNAc-DH is the best that can be done.

Lys_{HMM} and aLa_{HMM} were generated from sequences belonging to Chordates (and Arthropods in case of Lys_{HMM}) and hits of these profiles expectedly belong to the same phyla. HexNAc-DHs have not been identified in eukaryotes as yet, but Hex-DHs contain members from all domains. Nevertheless, hits of HexNAc-DH_{HMM} contain sequences from eukaryotes (plants, nematodes, arthropods) and virus, indicating that sequence divergence among HexNAc-DHs are taxa independent. However, this may not always be true. Experimentally characterized GDP-mannose

pyrophosphorylases from eukaryotes and bacteria show more similarity within than across domains. In such cases, profile HMMs generated using sequences from only one domain (e.g., eukaryotes) will not be sensitive to sequences from the other domain.

How essential is the supplementation of ROC curve with a bit score distribution plot for arriving at optimal thresholds? If an HMM represents a functional family that is distinctly different from other functional families in its primary structure, then an ROC curve should suffice. This will be the case if we build an HMM to represent the NDP-sugar 6-dehydrogenase family (Figure 2). However, if one were to have two HMMs each of which represents functional families that vary from each other only at a few homologous positions in the primary structure, then supplementing ROC curves with a bit score distribution plot will significantly improve the optimality of thresholds. This is the case with HexNAc-DH_{HMM} and Hex-DH_{HMM}, and with Lys_{HMM} and aLa_{HMM} (Table 1). The extent of sequence divergence that accounts for functional variation (substrate specificity in this case) is marginal necessitating the supplementation.

Case Studies 3 and 4 involve identifying homologs of SH3 and GT-A domains. Unlike as in Case Studies 1 and 2, ROC curve or any other existing annotation-based approaches cannot be utilized even for the identification of domain homologs as annotations of most of their hits do not include domain information. As there are no benchmark datasets to determine thresholds, there is a larger compromise with sensitivity. Pfam has subclassified Src-homology 3 (SH3) family into 37 subfamilies based on sequence similarity and assigned gathering thresholds (Punta et al. 2012). Scoring a dataset containing seed sequences of all the 37 subfamilies, we found overlapping hits for multiple profiles, suggesting the need for revisiting gathering thresholds. We find that thresholds for these families can be resolved by either (i) merging subfamilies that contain high scoring false positives from each other, (ii) using additional approaches such as motif search in combination with profile HMMs, or (iii) considering two thresholds, one with high specificity (stringent threshold) another with high sensitivity (relaxed threshold). There is no straightforward

way to assign optimal thresholds and a compromise between specificity or sensitivity is inevitable in the absence of benchmark datasets. Unlike SH3 domains that are involved in protein-protein interactions, GT-A fold (Case Study 4) is responsible for catalytic activity. However, like SH3 domain, it is often found as a part of a multidomain protein. Often, sequence-based approaches transfer molecular function annotation of one domain to the entire protein (Purushotham et al. 2020). Under these scenarios, it is important to define domain boundaries as a part of functional annotation. We used structurally characterized glycosyltransferases from GT2 family (that adopt GT-A fold) to illustrate an approach to demarcate domain boundaries. A manually curated dataset of experimentally characterized sequences classified under GT2 family in CAZy (Lombard et al. 2014) was used as benchmark dataset to identify HMM coverage threshold. Additionally, sequences belonging to other CAZy families adopting GT-A fold were used to identify a bit-score threshold. Thus, a combination of HMM coverage and bit-score threshold are required to identify GT-A domains.

How to find a bit score threshold that is optimal for specificity even if the search is poorly sensitive? One approach is the use of "negative" datasets. For example, to set threshold for an HMM that represents a KEGG Ortholog (KO) family (considered as the positive dataset), KofamKOALA uses proteins belonging to other KO families as negative datasets (Aramaki et al. 2020). The distribution of bit scores for proteins belonging to the positive and negative datasets are used to arrive at the threshold. HMMERCTTER follows a similar approach and identifies thresholds for sequence clusters within a superfamily by comparing inter-cluster scores (Pagnuco et al. 2018). The advantage of these two methods is that they do not require subject matter experts (SMEs) and hence are scalable (i.e., amenable for automation). These methods work well for HMMs that represent sequence families that share broad molecular function e.g., nucleotide sugar 6-dehydrogenases (as at Level 1 in Figure 2). The approach discussed in the

present study is effective even for relatively fine grained function annotation (as at Levels 2 and 3 in Figure 2) because it involves SMEs; obviously, this is not truly scalable.

AUTHOR CONTRIBUTIONS

PVB conceptualized research, JS performed analysis, RM contributed to lysozyme-C and GT2 family analysis, AK contributed to SH3 domains analysis, PVB and JS cowrote the paper.

ACKNOWLEDGEMENTS

We thank Toshi Mishra and Busi Rakesh for useful discussions. 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].

References

Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, Ogata H. 2020. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics. 36(7):2251–2252. doi:10.1093/bioinformatics/btz859.

Borg AJE, Beerens K, Pfeiffer M, Desmet T, Nidetzky B. 2020. Stereo-electronic control of reaction selectivity in short-chain dehydrogenases: Decarboxylation, epimerization, and dehydration. Curr Opin Chem Biol. 61:43–52. doi:10.1016/j.cbpa.2020.09.010.

Carpentier M, Chomilier J. 2019. Protein multiple alignments: sequence-based versus structure-based programs. Bioinformatics. 35(20):3970–3980. doi:10.1093/bioinformatics/btz236.

Churchill GA. 1989. Stochastic models for heterogeneous DNA sequences. Bull Math Biol. 51(1):79–94. doi:10.1007/BF02458837.

Durbin R, Eddy SR, Krogh A, Mitchison G. 1998. Section 5.2, Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. 1st ed. Cambridge University Press.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39(Web Server issue):W29-37. doi:10.1093/nar/gkr367.

Gloster TM. 2014. Advances in understanding glycosyltransferases from a structural perspective. Curr Opin Struct Biol. 28:131–141. doi:10.1016/j.sbi.2014.08.012.

Gough J, Karplus K, Hughey R, Chothia C. 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol. 313(4):903–919. doi:10.1006/jmbi.2001.5080.

Grobler JA, Rao KR, Pervaiz S, Brew K. 1994. Sequences of two highly divergent canine type c lysozymes: implications for the evolutionary origins of the lysozyme/alpha-lactalbumin superfamily. Arch Biochem Biophys. 313(2):360–366. doi:10.1006/abbi.1994.1399.

Grobler JA, Wang M, Pike AC, Brew K. 1994. Study by mutagenesis of the roles of two aromatic clusters of alpha-lactalbumin in aspects of its action in the lactose synthase system. J Biol Chem. 269(7):5106–5114.

Gruszczyk J, Fleurie A, Olivares-Illana V, Béchet E, Zanella-Cleon I, Moréra S, Meyer P, Pompidor G, Kahn R, Grangeasse C, et al. 2011. Structure analysis of the Staphylococcus aureus UDP-N-acetyl-mannosamine dehydrogenase Cap5O involved in capsular polysaccharide biosynthesis. J Biol Chem. 286(19):17112–17121. doi:10.1074/jbc.M110.216002.

Herrero MB, Mandal A, Digilio LC, Coonrod SA, Maier B, Herr JC. 2005. Mouse SLLP1, a sperm lysozyme-like protein involved in sperm-egg binding and fertilization. Dev Biol. 284(1):126–142. doi:10.1016/j.ydbio.2005.05.008.

Hubbard TJ, Park J. 1995. Fold recognition and ab initio structure predictions using hidden Markov models and beta-strand pair potentials. Proteins. 23(3):398–402. doi:10.1002/prot.340230313.

Irwin DM, Biegel JM, Stewart C-B. 2011. Evolution of the mammalian lysozyme gene family. BMC Evol Biol. 11:166. doi:10.1186/1471-2148-11-166.

Kalra S, Pradeep MA, Mohanty AK, Kaushik JK. 2016. Structural, Functional and Phylogenetic Analysis of Sperm Lysozyme-Like Proteins. PLoS One. 11(11):e0166321. doi:10.1371/journal.pone.0166321.

Kattke MD, Gosschalk JE, Martinez OE, Kumar G, Gale RT, Cascio D, Sawaya MR, Philips M, Brown ED, Clubb RT. 2019. Structure and mechanism of TagA, a novel membrane-associated glycosyltransferase that produces wall teichoic acids in pathogenic bacteria. PLoS Pathog. 15(4):e1007723. doi:10.1371/journal.ppat.1007723.

Krogh A, Brown M, Mian IS, Sjölander K, Haussler D. 1994. Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol. 235(5):1501–1531. doi:10.1006/jmbi.1994.1104.

Krogh A, Mian IS, Haussler D. 1994. A hidden Markov model that finds genes in E. coli DNA. Nucleic Acids Res. 22(22):4768–4778. doi:10.1093/nar/22.22.4768.

Kurochkina N, Guha U. 2013. SH3 domains: modules of protein-protein interactions. Biophys Rev. 5(1):29–39. doi:10.1007/s12551-012-0081-z.

Lander ES, Green P. 1987. Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci U S A. 84(8):2363–2367. doi:10.1073/pnas.84.8.2363.

Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. 2014. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 42(Database issue):D490-495. doi:10.1093/nar/gkt1178.

Malcolm BA, Rosenberg S, Corey MJ, Allen JS, de Baetselier A, Kirsch JF. 1989. Site-directed mutagenesis of the catalytic residues Asp-52 and Glu-35 of chicken egg white lysozyme. Proc Natl Acad Sci U S A. 86(1):133–137. doi:10.1073/pnas.86.1.133.

Pagnuco IA, Revuelta MV, Bondino HG, Brun M, Ten Have A. 2018. HMMER Cut-off Threshold Tool (HMMERCTTER): Supervised classification of superfamily protein sequences with a reliable cut-off threshold. PLoS One. 13(3):e0193757. doi:10.1371/journal.pone.0193757.

Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C. 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol. 284(4):1201–1210. doi:10.1006/jmbi.1998.2221.

Prager EM, Wilson AC. 1988. Ancient origin of lactalbumin from lysozyme: analysis of DNA and amino acid sequences. J Mol Evol. 27(4):326–335. doi:10.1007/BF02101195.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, et al. 2012. The Pfam protein families database. Nucleic Acids Res. 40(Database issue):D290–D301. doi:10.1093/nar/gkr1065.

Purushotham P, Ho R, Zimmer J. 2020. Architecture of a catalytically active homotrimeric plant cellulose synthase complex. Science. 369(6507):1089–1094. doi:10.1126/science.abb2978.

Qasba PK, Kumar S. 1997. Molecular divergence of lysozymes and alpha-lactalbumin. Crit Rev Biochem Mol Biol. 32(4):255–306. doi:10.3109/10409239709082574.

Rabiner LR, Levinson SE, Sondhi MM. 1983. On the Application of Vector Quantization and Hidden Markov Models to Speaker-Independent, Isolated Word Recognition. Bell System Technical Journal. 62(4):1075–1105. doi:https://doi.org/10.1002/j.1538-7305.1983.tb03115.x.

Ramakrishnan B, Qasba PK. 2001. Crystal structure of lactose synthase reveals a large conformational change in its catalytic component, the beta1,4-galactosyltransferase-I. J Mol Biol. 310(1):205–218. doi:10.1006/jmbi.2001.4757.

Rocha J, Popescu AO, Borges P, Mil-Homens D, Moreira LM, Sá-Correia I, Fialho AM, Frazão C. 2011. Structure of Burkholderia cepacia UDP-glucose dehydrogenase (UGD) BceC and role of Tyr10 in final hydrolysis of UGD thioester intermediate. J Bacteriol. 193(15):3978–3987. doi:10.1128/JB.01076-10.

Snook CF, Tipton PA, Beamer LJ. 2003. Crystal structure of GDP-mannose dehydrogenase: a key enzyme of alginate biosynthesis in P. aeruginosa. Biochemistry. 42(16):4658–4668. doi:10.1021/bi027328k.

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

Stojmirović A, Gertz EM, Altschul SF, Yu Y-K. 2008. The effectiveness of position- and composition-specific gap costs for protein similarity searches. Bioinformatics. 24(13):i15-23. doi:10.1093/bioinformatics/btn171.

UniProt Consortium. 2021. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 49(D1):D480–D489. doi:10.1093/nar/gkaa1100.

Wei J, Li S-J, Shi H, Wang H-Y, Rong C-T, Zhu P, Jin S-H, Liu J, Li J-Y. 2013. Characterisation of Lyzls in mice and antibacterial properties of human LYZL6. Asian J Androl. 15(6):824–830. doi:10.1038/aja.2013.93.

White JV, Stultz CM, Smith TF. 1994. Protein classification by stochastic modeling and optimal filtering of amino-acid sequences. Math Biosci. 119(1):35–75. doi:10.1016/0025-5564(94)90004-3.

Yang J, Xu Y, Chen CS. 1994. Hidden Markov model approach to skill learning and its application to telerobotics. IEEE Trans Rob Autom. 10(5):621–631. doi:10.1109/70.326567.

Zhang H, Zhu F, Yang T, Ding L, Zhou M, Li J, Haslam SM, Dell A, Erlandsen H, Wu H. 2014. The highly conserved domain of unknown function 1792 has a distinct glycosyltransferase fold. Nat Commun. 5:4339. doi:10.1038/ncomms5339.

Zhang X-Y, Yan Q-X, Guo X-Y, Chen C-R, Chen R-Q, Cai Z-M, Tang A-F. 2016. Expression profile of SPACA5/Spaca5 in spermatogenesis and transitional cell carcinoma of the bladder. Oncol Lett. 12(5):3731–3738. doi:10.3892/ol.2016.5164.

Zhao X, Creuzenet C, Bélanger M, Egbosimba E, Li J, Lam JS. 2000. WbpO, a UDP-N-acetyl-D-galactosamine dehydrogenase from Pseudomonas aeruginosa serotype O6. J Biol Chem. 275(43):33252–33259. doi:10.1074/jbc.M004191200.