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

Supplementary material containing Figures S1-S7, Table S1

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
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

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. In this study, we illustrate 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. We illustrate ways in which an optimal bit score can be arrived at using four Case Studies. These are 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.

INTRODUCTION

Hidden Markov Model (HMM) is a statistical modelling approach that has gained success across diverse fields including Biology (1–4). Profile HMMs have been used for the identification of sequence homologs (5), prediction of coding regions (6), fold recognition (7, 8), etc. Profile HMMs were introduced in sequence analysis as models to capture conserved positions among sequence variants of a functional family (5). Profile HMMs outperform pairwise BLAST in detecting remote homologs (9). This is because pairwise BLAST assigns equal importance to all positions of a sequence and does not have a mechanism to consider position specific sequence variations among homologs. PSI-BLAST performs iterative search over sequence databases and generates position specific scoring matrices (10). Profile HMMs, in addition, assign position dependent gap penalties and penalize gaps in conserved regions more than in variable regions (5) and are thus more robust than PSI-BLAST. Consequently, profile HMM databases such as Pfam (11) and SUPERFAMILY (12) are being used for functional annotation of new protein sequences arising out of whole genome and metagenome sequencing projects.

A critical step in implementing profile HMMs for detecting functional homologs is the identification of an optimal E-value or bit-score threshold. Optimization typically involves a trade-off between specificity and sensitivity. However, choosing 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. To achieve high-throughput, an automated process for setting bit score thresholds may be conceived but optimality of such thresholds cannot be guaranteed since every family of functional homologs is

unique. A customized workflow which takes into consideration the nature and amount of experimental data available for setting the bit score threshold 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

Protein sequences were taken from the UniProtKB database, release 2021_01 (13). Protein 3D structures were from the Protein Data Bank (14). Pairwise sequence comparisons were performed using locally installed BLAST and/or the BLAST server (15). Structural comparisons were made using PyMOL (16). Multiple sequence alignments were obtained from MAFFT command line application (17). Sequence motifs were identified using the MEME suite (18). Default values were used for all the parameters in all software and web servers unless stated otherwise.

The HMMER software suite (http://hmmer.org/) was used for generating profile HMMs and for scoring sequences. Profiles were generated using a dataset consisting of experimentally characterized proteins taken from the literature and reviewed sequences from Swiss-Prot (Figure 1). This type of dataset is henceforth referred to as Exp_dataset. Profiles for SH3 domains were generated from seed alignments downloaded from the Pfam database (Release 33.1) (11). Bit scores were used as thresholds rather than E-values, since bit scores are independent of database size. The Exp_dataset sequences were scored against the corresponding profile and the lowest bit score for Exp_dataset sequences is

denoted as T_{exp}. Entries in the TrEMBL database are electronically annotated using UniRule and Association-Rule-Based Annotator (ARBA) (13). These were used to generate Receiver Operator Characteristic (ROC) curves, as described in more detail elsewhere (19). 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.

Position-specific amino acid frequency matrices were computed from multiple sequence alignments (MSA) 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 region 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

<u>Generating Profile HMMs</u>: Nucleotide sugar 6-dehydrogenases include HexNAc-DH and Hex-DH (Figure 2). They all belong to the Short-chain Dehydrogenase Reductase (SDR) superfamily (20). These enzymes catalyze the conversion of hexose / HexNAc to the corresponding uronic acid derivative. In order to identify proteins belonging to these two functional families in completely sequenced

genomes and metagenomes, two profiles viz., HexNAc-DH_{HMM} and Hex-DH_{HMM} were built (Figure 1). T_{exp} was found to be 536 and 250 bits for HexNAc-DH_{HMM} and Hex-DH_{HMM}, respectively (Table 1).

ROC curves: The TrEMBL database has nearly 17,000 sequences that are annotated as a HexNAc-DH. Scoring TrEMBL using HexNAc-DH_{HMM} with T_{exp} as the threshold suggested that the T_{exp} has poor sensitivity (Figure 1). An ROC curve was generated to improve the sensitivity of the profile (Figure 3A). Based on this ROC curve, the threshold was set to 256 bits (T_{ROC}); with this threshold, there was a marked increase in sensitivity. However, there is a concomitant increase in false positives also (Figure 1).

TrEMBL was scored using Hex-DH_{HMM} also with its T_{exp} as the threshold. 81% of the total ~53000 sequences annotated as Hex-DH 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 3B).

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. Can the threshold be fine-tuned to improve specificity? Towards this, TrEMBL was scored using both HexNAc-DH_{HMM} and Hex-DH_{HMM} using default values suggested by HMMER as thresholds. Bit scores of hits that are common to the two profiles were plotted against each other (Figure 4A). It is observed that proteins which are “high-scoring” for one profile are “low-scoring” against the other. This scatter 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). As can be expected, this revision resulted in a marginal increase in specificity and decrease in sensitivity. Barring a few, false positives of HexNAc-DH_{HMM} score higher than they score 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?

Residue conservation analysis: To analyze the optimality of the threshold, conservation of active site residues among hits was examined in both families of 6-dehydrogenases. In HexNAc-DHs, the cofactor binding motif GxGxxG, conserved cysteine, arginine and lysine (Lys204, Arg211 and Cys258, *Staphylococcus aureus* UDP-ManNAc 6-dehydrogenase numbering (21) 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) (22, 23) are conserved in ~95% of the hits. The bit scores of the 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. Nevertheless, the number of false positives based on residue conservation are very low, suggesting that the threshold derived from aforementioned approaches is reliable. Thus, while profile HMM is effective in identifying close homologs, combining it 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 (21, 23); 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 (19). However, it was challenging to resolve the substrate specificity among HexNAc-DHs because (i) some of these enzymes can utilize more than one substrate (24) 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

Lysozymes: Lysozymes are muramidases that lyse peptidoglycan complexes of bacteria. Based on primary and tertiary structure comparisons, lysozymes have been divided into different families e.g., c-type (chicken type; lysozyme-C family), i-type (invertebrate type), g-type (goose type), phage T4 type, bacterial and plant lysozymes (25). Only lysozymes belonging to the lysozyme-C family are considered in this study, unless specified otherwise.

Sequence and structure homologs with mutually exclusive functions: Lysozyme-C family is named after hen egg white lysozyme, its prototypical member. This

family includes lysozymes, lysozyme-like proteins and alpha-lactalbumins (26–29). Some lysozymes bind calcium [Ca(+)-Lys] whereas others do not [Ca(-)-Lys]. 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 (30). Phylogenetic analyses have indicated that alpha-lactalbumins and lysozymes diverged from a common ancestor, a calcium binding protein (28, 29) (Figure S1). The vestiges of divergence are evident in lysozymes that retain calcium-binding ability. Lysozyme-like proteins are suggested to have diverged much later during the evolution of vertebrates (28) but there is a paucity of experimental data on these proteins.

Lysozyme-C family members exhibit high sequence and 3D structure similarities (31) (Figure S2) despite of functional divergence. alpha-Lactalbumins (27) and lysozyme-like proteins (28) 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.

<u>Profile HMMs could not distinguish calcium-binding lysozymes from those which do not bind calcium</u>: Both Ca(+)-Lys and Ca(-)-Lys retain catalytic activity. The ability to bind calcium is imparted by a 7-residue Asp-rich motif (32). Two profiles were generated (Table 1) 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 (Table 1), indicating that HMMs cannot distinguish Ca(+) and Ca(-)-Lys from each other based on available experimental data. Consequently, both Ca(+)-Lys and Ca(-)-Lys were combined to form a single profile viz., Lys_{HMM}.

Setting thresholds: Two profiles viz., Lys_{HMM} and aLa_{HMM} were generated for lysozyme and alpha-lactalbumins (Figure 1; Table 1). T_{exp} for these two profiles are 168 and 184 bits, respectively (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 139 bits; similarly, the highest score of a lysozyme against aLa_{HMM} is 138 bits. An ROC curve could not be plotted for Lys_{HMM} as there were no false positives scoring >139 bits among its hits from TrEMBL. Therefore, we used bit score comparison approach to set thresholds for both Lys_{HMM} and aLa_{HMM} (Figure 4B). Accordingly, T_{Final} was set to 140 bits for Lys_{HMM} and 150 bits for aLa_{HMM} (Table 1).

Analysis of annotations of TrEMBL hits to Lys_{HMM}: Using T_{Final} (Table 1) as the threshold, 1706 hits were obtained from TrEMBL for Lys_{HMM}. Out of these, very few are annotated as lysozyme-C while a majority have incomplete annotations viz., “lysozyme”, “lysozyme-like”, “alpha-lactalbumin/lysozyme”, etc., 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 (33)). Scores of such sequences fall in the range 140-236 bits. Several proteins are annotated as ‘sperm acrosome membrane-associated protein’ and these are described in literature as lysozyme-like proteins, some of which have lost their bacteriolytic activity due to mutations in catalytic site (34–36). As observed earlier, profile HMMs are insensitive to point mutations even in this protein family and thus, cannot differentiate catalytically active proteins from inactive homologs. 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 hits to the aLA_{HMM}, of which one has no molecular function annotation (Figure 1). 105 sequences are annotated as 'Lactose synthase B' and hence are true positives [note: alpha-lactalbumin is also denoted as domain B of lactose synthase complex (https://www.brenda-enzymes.org/enzyme.php?ecno=2.4.1.22#:~:text=The%20enzyme%20is%20a%20complex,1.90%20N%2Dacetyllactosamine%20synthase)]. Sequences of true positives (i.e., those hits which are annotated as alpha-lactalbumin) were analysed for conservation of functionally important residues viz., Phe50, His51, Gln138 and Trp137 (bovine alpha-lactalbumin numbering (37)). Phe50 and His51 are conserved among all hits but Gln136 and Trp137, residues implicated in binding to galactosyltransferase (37), 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 fold

A highly divergent sequence family: Src homology 3 (SH3) domains are ~60-residue long structural elements that mediate protein-protein interactions by recognizing proline rich motifs (38, 39). They are 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 (11) (Supplementary_dataset.xlsx: SH3 families). SH3 domains are a part of multi-domain proteins that mediate diverse biological processes (40). As a result of this, 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 (11). Herein, we examine the domain gathering thresholds of SH3 domain profile HMMs (as of May 2021).

Do seed alignment sequences meet thresholds of >1 profile HMM? 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 (11). 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 (11). In fact, increasing the gathering threshold by ~10 bits improved the specificity of some of the profiles (discussed below).

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 six families, multiple hits with scores comparable to that of seed sequences were obtained (Figure 5).

Resolving false positives: For this purpose, one can either employ residue/motif conservation analyses over and above profile HMMs, or combine their seed sequences into a single profile and redefine their annotation (Figure 5). We will describe the motif conservation approach by considering SH3_1 and SH3_2 families as an illustrative example. Nearly 80% of SH3_1 seed sequences are hits to SH3_2 profile (Figure S4). Visual inspection of multiple sequence alignments of seed sequences of these two families reveals a 6 amino acid extension in SH3_2 that is absent in SH3_1 (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.

Reclassification of a seed sequence: A sequence from GW family (Q5KV47_GEOKA/411-472) scores 25.8 bits against GW profile but 38.4 and 24.2

bits against SH3_3 and SH3_4 profiles, respectively. The 62 aa sequence scores with higher sequence coverage against SH3_3 and SH3_4 profiles. The seed alignment of GW family consisting of this sequence depicts long gaps, raising the possibility of reclassifying this sequence into SH3_3 or SH3_4.

Case Study 4: GT-A fold of glycosyltransferases

GT2 family: The CAZy database describes families of structurally-related catalytic and carbohydrate-binding modules (or functional domains) of enzymes that degrade, modify, or create glycosidic bonds (41). Glycosyltransferases (GTs), one of the enzyme classes covered in this database, 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.

GT-A fold: 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 (42–44). Such a conservation of fold despite of sequence and functional differences is not surprising since it has long been known that 3D structures are more conserved than sequences. 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 S1). 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 structure 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 (45).

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) 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).

<u>GT-A fold domain in other GT families of CAZy</u>: 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.

Discussion

Profile HMMs are powerful tools to detect remote homology in sequence databases and newly 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 of a polypeptide that may be vestiges of divergent evolution. 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 some of the 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 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 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. 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.

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 (11). 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 (46). 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 (41) 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.

Several *in-silico* approaches continue to be developed by researchers across the globe to address challenges in molecular function annotation among different protein families. Improved annotations resulting from such efforts are typically not integrated into publicly accessible protein sequence databases. A unified platform allowing researchers to share annotations will contribute positively towards improving electronic annotations.

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

1. J. Yang, Y. Xu, C. S. Chen, Hidden Markov model approach to skill learning and its application to telerobotics. *IEEE Trans Rob Autom* **10**, 621–631 (1994).

2. L. R. Rabiner, S. E. Levinson, M. M. Sondhi, On the Application of Vector Quantization and Hidden Markov Models to Speaker-Independent, Isolated Word Recognition. *Bell System Technical Journal* **62**, 1075–1105 (1983).

3. E. S. Lander, P. Green, Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci U S A* **84**, 2363–2367 (1987).

4. G. A. Churchill, Stochastic models for heterogeneous DNA sequences. *Bull Math Biol* **51**, 79–94 (1989).

5. A. Krogh, M. Brown, I. S. Mian, K. Sjölander, D. Haussler, Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol* **235**, 1501–1531 (1994).

6. A. Krogh, I. S. Mian, D. Haussler, A hidden Markov model that finds genes in E. coli DNA. *Nucleic Acids Res* **22**, 4768–4778 (1994).

7. J. V. White, C. M. Stultz, T. F. Smith, Protein classification by stochastic modeling and optimal filtering of amino-acid sequences. *Math Biosci* **119**, 35–75 (1994).

8. T. J. Hubbard, J. Park, Fold recognition and ab initio structure predictions using hidden Markov models and beta-strand pair potentials. *Proteins* **23**, 398–402 (1995).

9. J. Park, *et al.*, Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J Mol Biol* **284**, 1201–1210 (1998).

10. S. F. Altschul, *et al.*, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402 (1997).

11. M. Punta, *et al.*, The Pfam protein families database. *Nucleic Acids Res* **40**, D290–D301 (2012).

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

13. UniProt Consortium, UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* **49**, D480–D489 (2021).

14. H. M. Berman, *et al.*, The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

15. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

16. , PyMOL | pymol.org (May 18, 2021).

17. K. Katoh, D. M. Standley, MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **30**, 772–780 (2013).

18. T. L. Bailey, C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**, 28–36 (1994).

19. J. Srivastava, P. Sunthar, P. V. Balaji, The glycan alphabet is not universal: a hypothesis. *Microb Genom* **6** (2020).

20. A. J. E. Borg, K. Beerens, M. Pfeiffer, T. Desmet, B. Nidetzky, Stereo-electronic control of reaction selectivity in short-chain dehydrogenases: Decarboxylation, epimerization, and dehydration. *Curr Opin Chem Biol* **61**, 43–52 (2020).

21. J. Gruszczyk, *et al.*, Structure analysis of the Staphylococcus aureus UDP-N-acetyl-mannosamine dehydrogenase Cap5O involved in capsular polysaccharide biosynthesis. *J. Biol. Chem.* **286**, 17112–17121 (2011).

22. C. F. Snook, P. A. Tipton, L. J. Beamer, Crystal structure of GDP-mannose dehydrogenase: a key enzyme of alginate biosynthesis in P. aeruginosa. *Biochemistry* **42**, 4658–4668 (2003).

23. J. Rocha, *et al.*, Structure of Burkholderia cepacia UDP-glucose dehydrogenase (UGD) BceC and role of Tyr10 in final hydrolysis of UGD thioester intermediate. *J. Bacteriol.* **193**, 3978–3987 (2011).

24. X. Zhao, *et al.*, WbpO, a UDP-N-acetyl-D-galactosamine dehydrogenase from Pseudomonas aeruginosa serotype O6. *J. Biol. Chem.* **275**, 33252–33259 (2000).

25. P. Jollès, J. Jollès, What's new in lysozyme research? Always a model system, today as yesterday. *Mol Cell Biochem* **63**, 165–189 (1984).

26. J. A. Grobler, K. R. Rao, S. Pervaiz, K. Brew, Sequences of two highly divergent canine type c lysozymes: implications for the evolutionary origins of the lysozyme/alpha-lactalbumin superfamily. *Arch Biochem Biophys* **313**, 360–366 (1994).

27. E. M. Prager, A. C. Wilson, Ancient origin of lactalbumin from lysozyme: analysis of DNA and amino acid sequences. *J Mol Evol* **27**, 326–335 (1988).

28. D. M. Irwin, J. M. Biegel, C.-B. Stewart, Evolution of the mammalian lysozyme gene family. *BMC Evol Biol* **11**, 166 (2011).

29. S. Kalra, M. A. Pradeep, A. K. Mohanty, J. K. Kaushik, Structural, Functional and Phylogenetic Analysis of Sperm Lysozyme-Like Proteins. *PLoS One* **11**, e0166321 (2016).

30. B. Ramakrishnan, P. K. Qasba, Crystal structure of lactose synthase reveals a large conformational change in its catalytic component, the beta1,4-galactosyltransferase-I. *J Mol Biol* **310**, 205–218 (2001).

31. P. K. Qasba, S. Kumar, Molecular divergence of lysozymes and alpha-lactalbumin. *Crit Rev Biochem Mol Biol* **32**, 255–306 (1997).

32. H. Tsuge, *et al.*, Crystallographic studies of a calcium binding lysozyme from equine milk at 2.5 A resolution. *J Biochem* **111**, 141–143 (1992).

33. B. A. Malcolm, *et al.*, 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**, 133–137 (1989).

34. X.-Y. Zhang, *et al.*, Expression profile of SPACA5/Spaca5 in spermatogenesis and transitional cell carcinoma of the bladder. *Oncol Lett* **12**, 3731–3738 (2016).

35. M. B. Herrero, *et al.*, Mouse SLLP1, a sperm lysozyme-like protein involved in sperm-egg binding and fertilization. *Dev Biol* **284**, 126–142 (2005).

36. J. Wei, *et al.*, Characterisation of Lyzls in mice and antibacterial properties of human LYZL6. *Asian J Androl* **15**, 824–830 (2013).

37. J. A. Grobler, M. Wang, A. C. Pike, K. Brew, 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**, 5106–5114 (1994).

38. B. J. Mayer, M. J. Eck, SH3 domains. Minding your p's and q's. *Curr Biol* **5**, 364–367 (1995).

39. K. Saksela, P. Permi, SH3 domain ligand binding: What's the consensus and where's the specificity? *FEBS Lett* **586**, 2609–2614 (2012).

40. N. Kurochkina, U. Guha, SH3 domains: modules of protein-protein interactions. *Biophys Rev* **5**, 29–39 (2013).

41. V. Lombard, H. Golaconda Ramulu, E. Drula, P. M. Coutinho, B. Henrissat, The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490-495 (2014).

42. T. M. Gloster, Advances in understanding glycosyltransferases from a structural perspective. *Curr Opin Struct Biol* **28**, 131–141 (2014).

43. H. Zhang, *et al.*, The highly conserved domain of unknown function 1792 has a distinct glycosyltransferase fold. *Nat Commun* **5**, 4339 (2014).

44. M. D. Kattke, *et al.*, Structure and mechanism of TagA, a novel membrane-associated glycosyltransferase that produces wall teichoic acids in pathogenic bacteria. *PLoS Pathog* **15**, e1007723 (2019).

45. M. Carpentier, J. Chomilier, Protein multiple alignments: sequence-based versus structure-based programs. *Bioinformatics* **35**, 3970–3980 (2019).

46. P. Purushotham, R. Ho, J. Zimmer, Architecture of a catalytically active homotrimeric plant cellulose synthase complex. *Science* **369**, 1089–1094 (2020).

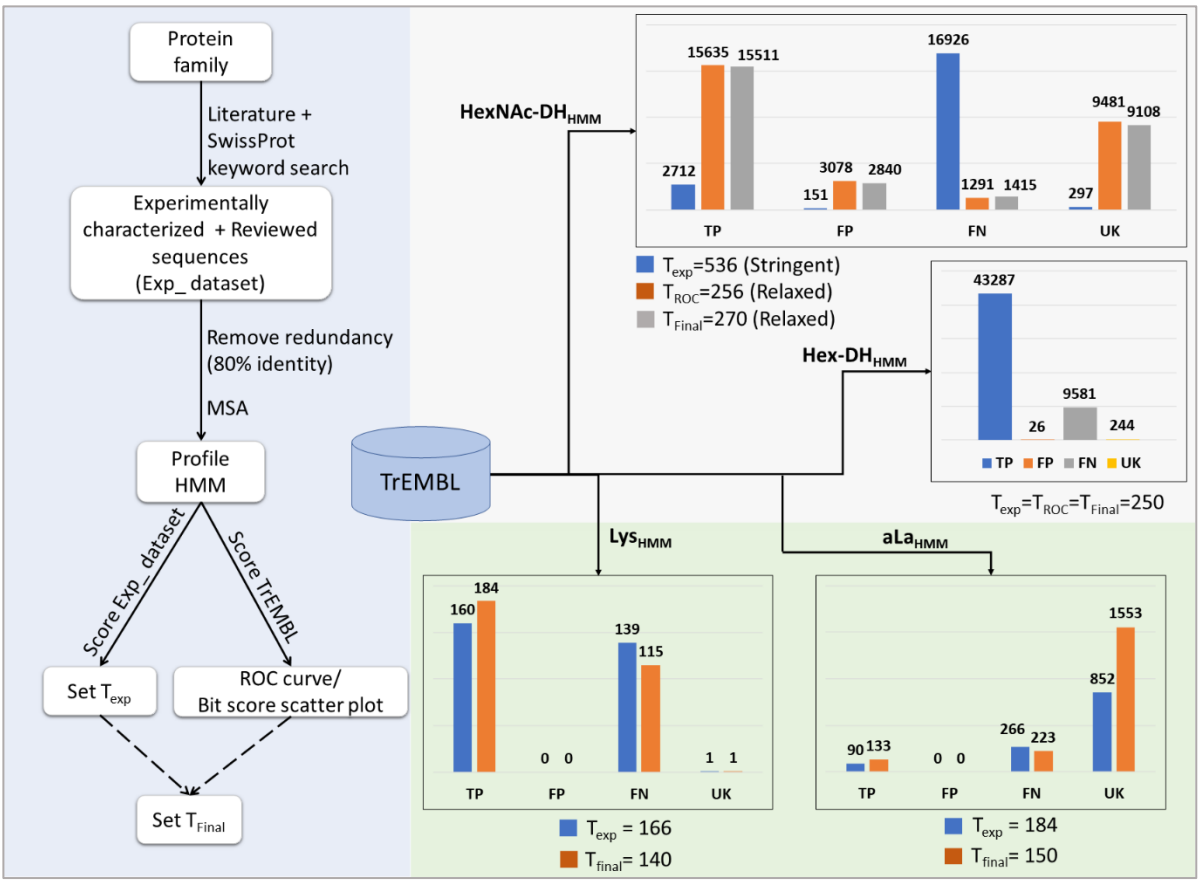


Figure 1: Procedure used to derive Profile HMMs and analyse hits for functional families in Case Studies 1 and 2. The final threshold, T_{Final}, can be the same as T_{exp} or T_{ROC}, or is set based on a bit score scatter plot. Details of Profile HMMs are given in Table 1. Details of proteins that constituted Exp_dataset are given as Supplementary_dataset.xlsx.
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.

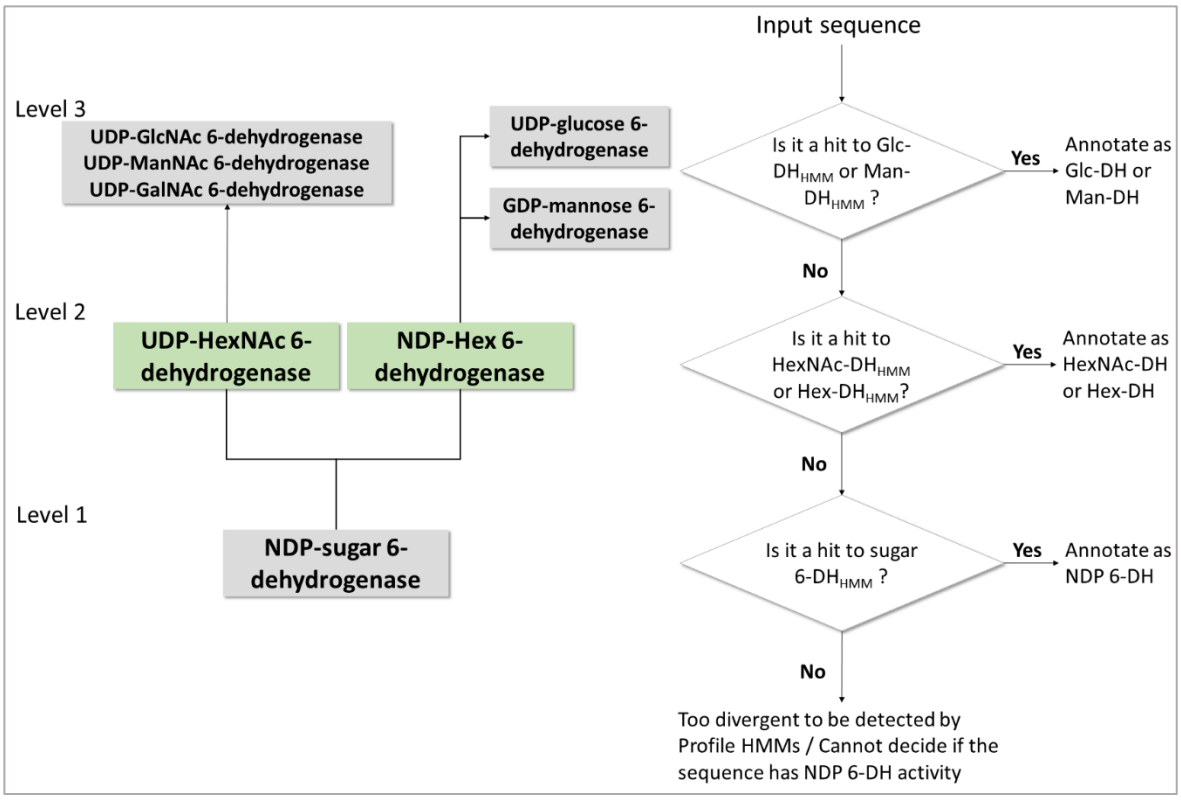


Figure 2: Categorization of nucleotide sugar 6-dehydrogenase family proteins into different levels based on substrate specificity. Assignment of substrate specificity is generic in Level 1 and increasingly specific in Levels 2 and 3. Separate profile HMMs could not be generated for the three HexNAc 6-DHs 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.

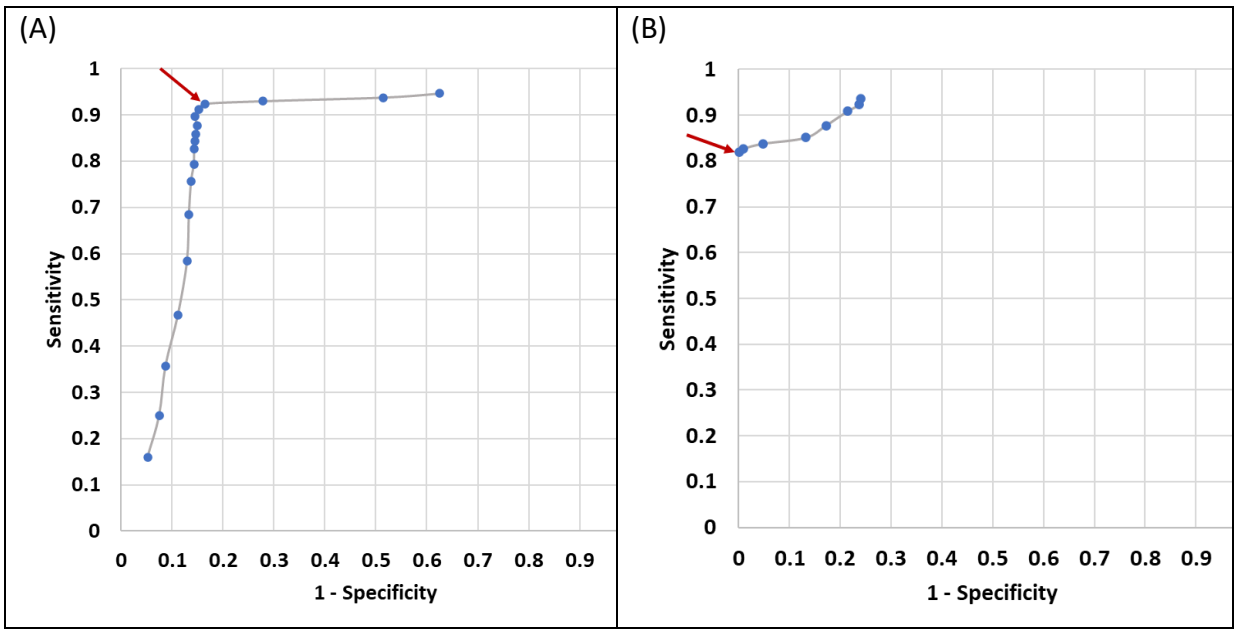


Figure 3: ROC curves for (A) HexNAc-DH_{HMM} and (B) Hex-DH_{HMM} obtained by scoring the TrEMBL database. Data points have been merely joined by a grey line. Arrow points to the data point corresponding to the bit score threshold determined from the curve viz., T_{ROC}.

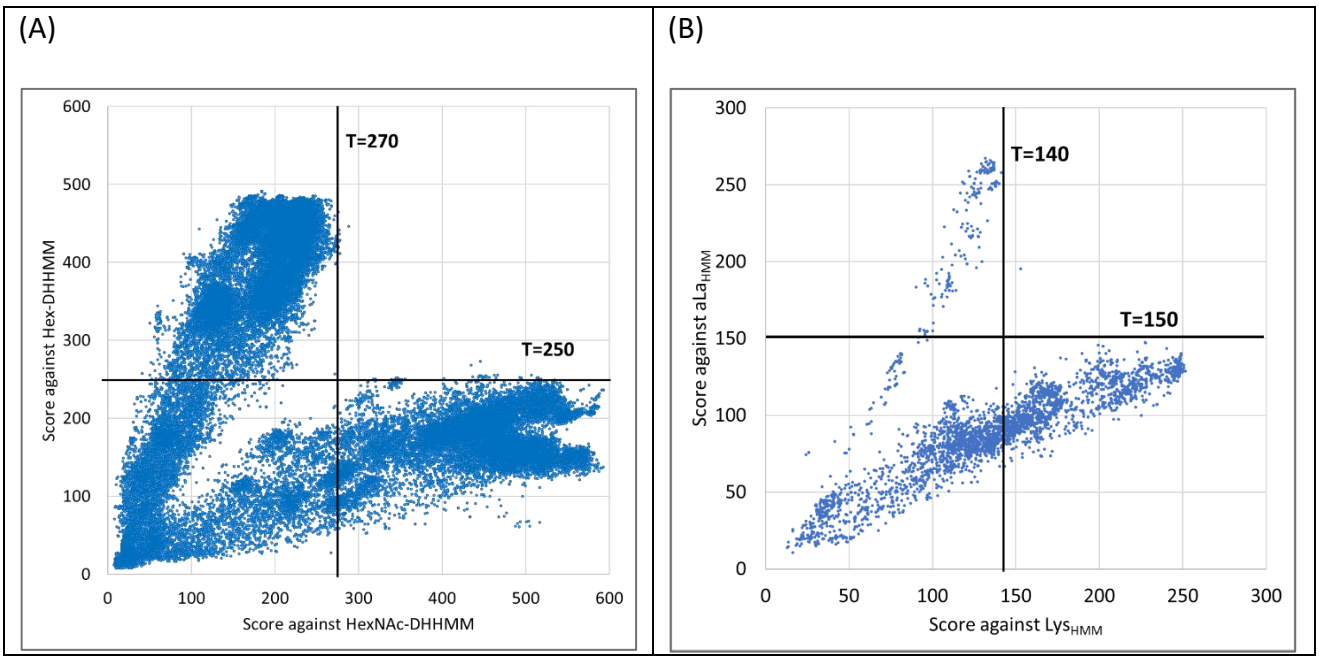
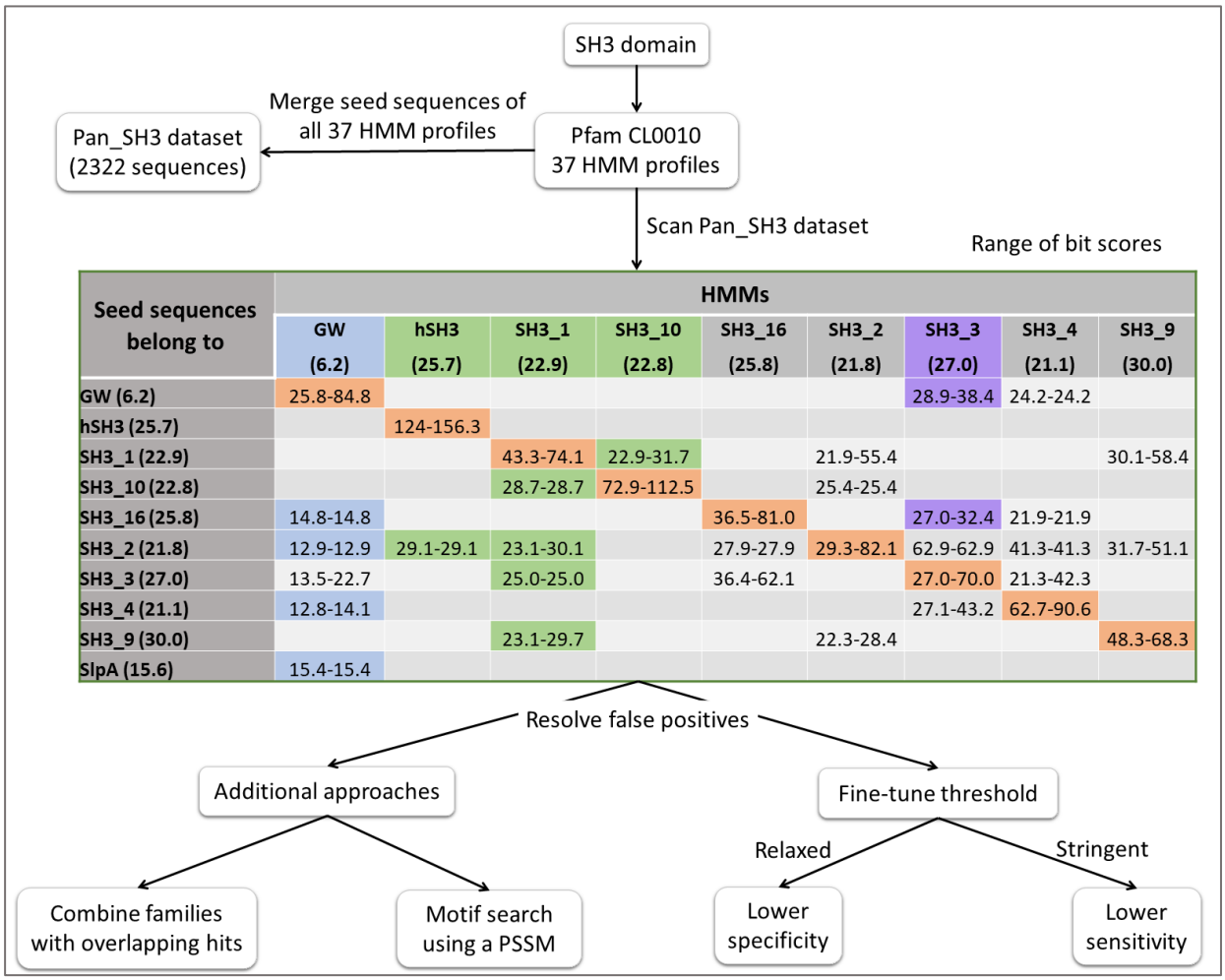


Figure 4: (A) Bit score comparison of TrEMBL entries that are hits to both Hex-DH$_{HMM}$ and HexNAc-DH$_{HMM}$. (B) Bit score comparison of common hits from TrEMBL against Lys$_{HMM}$ and aLa$_{HMM}$. Threshold was set to 140 and 150, respectively.



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 (62 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.

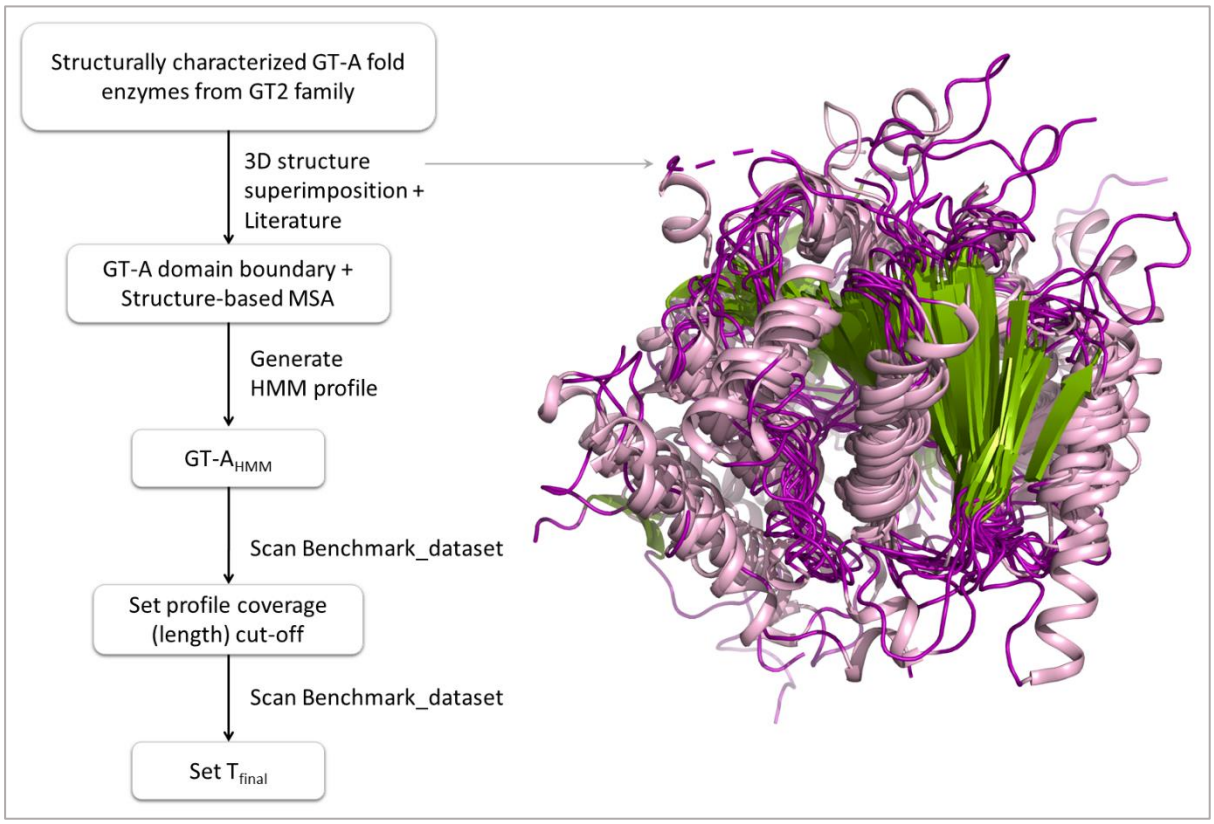


Figure 6: Steps followed to determine bit score threshold and HMM coverage cut-off for GT-A_{HMM}. PDB Ids and related information for proteins used for superposition are given in Table S1, Supplementary_dataset.xlsx:GT2 structures.

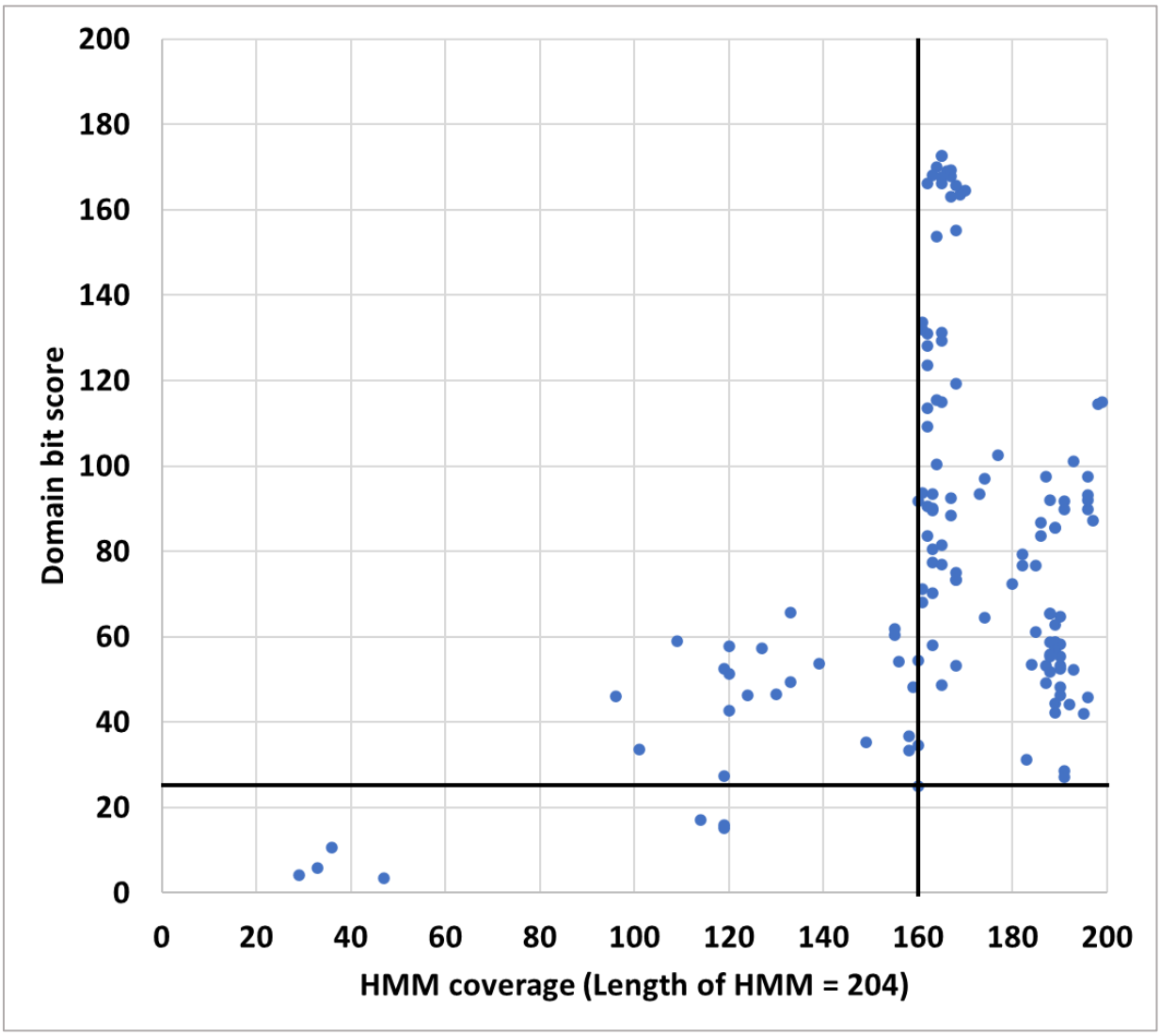


Figure 7: Bit scores of 133 sequences from GT2 benchmark dataset plotted against the corresponding HMM coverage (Supplementary_dataset.xlsx: GT-A_{HMM} hits). The horizontal line represents bit score cut off (=25 bits) and the vertical line represents HMM coverage cut-off (=160).