frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 02 August 2021
doi: 10.3389/fbioe.2021.701120

Functional Classification of Super-Large Families of Enzymes Based on Substrate Binding Pocket Residues for Biocatalysis and Enzyme Engineering Applications

Fernanda L. Sirota[1], *Sebastian Maurer-Stroh*[1,2], *Zhi Li*[3], *Frank Eisenhaber*[1,4,5]* *and Birgit Eisenhaber*[1,4]*

[1]*Bioinformatics Institute (BII), Agency for Science Technology and Research (A*STAR), Singapore, Singapore,* [2]*Department of Biological Sciences, National University of Singapore, Singapore, Singapore,* [3]*Department of Chemical and Biomolecular Engineering, National University of Singapore, Singapore, Singapore,* [4]*Genome Institute of Singapore (GIS), Agency for Science, Technology and Research (A*STAR), Singapore, Singapore,* [5]*School of Biological Sciences, Nanyang Technological University, Singapore, Singapore*

Large enzyme families such as the groups of zinc-dependent alcohol dehydrogenases (ADHs), long chain alcohol oxidases (AOxs) or amine dehydrogenases (AmDHs) with, sometimes, more than one million sequences in the non-redundant protein database and hundreds of experimentally characterized enzymes are excellent cases for protein engineering efforts aimed at refining and modifying substrate specificity. Yet, the backside of this wealth of information is that it becomes technically difficult to rationally select optimal sequence targets as well as sequence positions for mutagenesis studies. In all three cases, we approach the problem by starting with a group of experimentally well studied family members (including those with available 3D structures) and creating a structure-guided multiple sequence alignment and a modified phylogenetic tree (aka binding site tree) based just on a selection of potential substrate binding residue positions derived from experimental information (not from the full-length sequence alignment). Hereupon, the remaining, mostly uncharacterized enzyme sequences can be mapped; as a trend, sequence grouping in the tree branches follows substrate specificity. We show that this information can be used in the target selection for protein engineering work to narrow down to single suitable sequences and just a few relevant candidate positions for directed evolution towards activity for desired organic compound substrates. We also demonstrate how to find the closest thermophile example in the dataset if the engineering is aimed at achieving most robust enzymes.

Keywords: zinc-dependent alcohol dehydrogenase, ADH, sequence alignment conflict, substrate binding pocket, substrate specificity, enzyme engineering

OPEN ACCESS

Edited by:
Michele Galluccio, University of Calabria, Italy

Reviewed by:
Hector Riveros-Rosas, Universidad Nacional Autónoma de México, Mexico
Rajendran Velmurugan, Chulalongkorn University, Thailand

***Correspondence:**
Birgit Eisenhaber
birgite@bii.a-star.edu.sg
Frank Eisenhaber
franke@bii.a-star.edu.sg

Specialty section:
This article was submitted to Synthetic Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *27 April 2021*
Accepted: *12 July 2021*
Published: *02 August 2021*

Citation:
Sirota FL, Maurer-Stroh S, Li Z, Eisenhaber F and Eisenhaber B (2021) Functional Classification of Super-Large Families of Enzymes Based on Substrate Binding Pocket Residues for Biocatalysis and Enzyme Engineering Applications. Front. Bioeng. Biotechnol. 9:701120. doi: 10.3389/fbioe.2021.701120

INTRODUCTION

Biocatalysis has gained importance both through methodological advances like enzyme engineering and directed evolution of enzymes towards new substrates (Arnold, 2019) as well as trends towards green chemical manufacturing (Dunn, 2012). Several large enzyme families are prominent candidates for biotechnology applications including enzyme engineering for certain substrate specificities because of the wide range of organic chemistry transformations that can be supported with them. Zinc-dependent alcohol dehydrogenases (ADHs; enzyme classification EC 1.1.1.1), long chain alcohol oxidases (AOxs; enzyme classification 1.1.3.20) and amino dehydrogenases (AmDHs, enzyme classification 1.4.1.20) are popular examples.

For example, zinc-dependent ADHs are part of a very large family of enzymes (enzyme classification 1.1.1.*) catalyzing the reversible oxidation of diverse alcohols to aldehydes or ketones with the associated reduction of nicotinamide adenine dinucleotide (NAD+) or chemically similar co-factors. The degree of substrate specificity varies to the extent that even non-catalytic examples are known. Whereas some ADHs process just a narrow compound list, others have a large hydrophobic pocket that can handle a wide variety of small molecules but also much larger hydroxylated hydrophobic chains, cyclical or steroidal molecules such as bile alcohols, retinol, derivatives of epinephrine, serotonin, dopamine and leukotriene catabolism (Petruszko, 1979; Riveros-Rosas et al., 1997; Hoog et al., 2003; Riveros-Rosas et al., 2003; Persson et al., 2008). Even aldehyde oxidation to acids by dismutation is possible in some cases (Henehan and Oppenheimer, 1993). ADHs are extremely widespread; they have been identified in organisms ranging from prokaryotes to higher eukaryotes and have been studied for decades, in particular, the ones belonging to *Saccharomyces cerevisiae* (de Smidt et al., 2008), due to their importance and historical impact in fermentation.

The advantages of enzymatic synthesis become especially obvious in the case of region- and stereo-selective organic chemistry targets as governing the reaction towards pure yield is difficult and costly, if not practically impossible without biotechnological methods (Wu et al., 2021). For example, ethyl (R)-4-chloro-3-hydroxybutanoate ((R)-ECHB) is a chiral molecule applicable for the synthesis of biologically important compounds such as (R)-carnitine, (R)-4-hydroxy-2-pyrrolidone, (R)-4-amino-3hydroxy-butyric acid, etc. It can be synthesized with high yield and purity by using (S)-selective secondary alcohol dehydrogenase produced by *Candida parapsilosis* (CpSADH) overexpressed in a bacterial system (Yamamoto et al., 1995; Yamamoto et al., 1999; Yamamoto et al., 2002). A version of CpSADH with W296A mutation engineered from the enantioselective form creates an ambidextrous enzyme that can be widely used to oxidize alcohols and to feed them into cascade reactions with co-factor recycling, for example for the production of enantiopure amines (Tian and Li, 2020).

Similarly, the wealth of available protein sequences representing long chain alcohol oxidases (AOXs) and amine dehydrogenases (AmDHs) make them as attractive for substrate- and product-specific engineering as ADHs. Alcohol oxidases (alcohol:O_2 oxidoreductases; EC 1.1.3. x) carry a flavin coenzyme and catalyze the oxidation of alcohols to carbonyl with concomitant production of H_2O_2 (Goswami et al., 2013). These enzyme are widely used in biosensors and for industrial production of a wide range of carbonyl compounds (Thungon et al., 2017).

AmDHs (EC 1.4.99.3) are associated with a tryptophan tryptophylquinone (TTQ) cofactor and are known for the interconversion of ketones (with participation of ammonia) and enantiomerically pure amines (Knaus et al., 2017). Enzyme engineering of AmDHs for bio-catalytic chiral amine synthesis is widely described in the literature (Kohls et al., 2014; Tseliou et al., 2019a; Tseliou et al., 2019b).

At the start of such an enzyme engineering project, several candidate genes encoding a protein with substrate specificity close to the desired one need to be selected from the sequence databases. In addition, requirements with regard to enzyme thermostability (sourcing from thermophile organisms), optimal pH or salt concentration, etc. might be further constraints. If the selection has to be made from a pool of many thousands or millions of potentially suitable sequences as in the case of ADHs, AOxs or AmDHs, this is a daunting task.

In this work, we suggest a methodology that can shrink the set of candidate enzyme sequences to manually manageable sets and, yet, retain the optimal targets with high likelihood. The main idea exploits the fact that substrate specificity is not so much determined by the total sequence of the enzyme but, to a large extent, only by the residues that make up the surface of the catalytic and binding cavities. These binding site residues ultimately play the main role in determining the enzymes' substrate specificity and they are the preferred sites for directed evolution or site-directed mutations in the enzyme engineering process. Once the list and the identities of these residues are known, the total sequence set can then be sub-clustered with regard to similarities among this residue list. The tools for constructing phylogenetic trees can be used for this purpose; thus, enzymes with similar binding cavity surface will tend to be grouped into the same branch of the tree.

Clearly, the trees obtained this way are just "binding cavity similarity trees" or "binding site trees" and they do not necessarily reflect the real evolution of the enzyme families. It is more an approach towards sensible hypothesis construction and candidate sequence selection. First, the associated bootstrap values will not be informative given the short alignment is used for tree construction. Second, as the introduction of an active site requires only a few mutations and is much less expensive in evolutionary terms than the change of a fold, the emergence of active and binding sites with similar/identical specificity can happen independently in different branches of the evolutionary tree. Similarly, one and the same evolutionary branch can develop a variety of specificities (for example, in the case of the BindGPILA domain (Eisenhaber et al., 2018)).

In order to facilitate use of ADHs, AOxs and AmDHs for biocatalysis of new substrates, we show exemplarily how to select enzyme sequences most suited for engineering towards the new target substrate. We start from a list of already characterized enzymes belonging to these families and extract the set of residues

known to be involved in substrate binding. Then, we classify all sequence members of the respective enzyme family according to their different substrate specificity. To achieve this, we apply similarity tree-generating tools (applied onto key binding pocket residues) used in phylogenetic studies together with a few other bioinformatics methods.

METHODS

We describe the methodological approach in great detail for the group of zinc-dependent alcohol dehydrogenases (ADHs). The processing of protein sequences in the cases of AOxs and AmDHs involves the same steps and tools (see below for further detail).

Seed Alignment for the Alcohol Dehydrogenase Family

Four for each taxonomic group, well studied fungal (*C. parapsilosis*) and human sequences with 3D structures available (PDB (Burley et al., 2021) entries: 1U3U, 1U3V, 1U3W, 1U3T (Gibbons and Hurley, 2004), 3WLF, 3WLE, 3WNQ (Wang et al., 2014) and 4C4O (Man et al., 2014)) were selected for seeding the sequence family. These were aligned with MAFFT using Linsi parameters (Katoh and Standley, 2013) and manually curated in Jalview (Waterhouse et al., 2009) for structural equivalency of aligned residues taking into account information from structural alignment with MUSTANG (Konagurthu et al., 2006) in YASARA (Krieger and Vriend, 2014).

Alignment of Alcohol Dehydrogenases With Known Substrates

263 ADH-related sequences were retrieved from UniProt (UniProt Consortium, 2019) with filters for 1) enzyme classification EC 1.1.1.*, 2) belonging to the zinc containing alcohol dehydrogenase family, 3) protein sequence length ranging from 250 to 600 residues and 4) annotation for substrate catalytic activity with experimental evidence. The length restriction was to ensure a good coverage of entries with the domain architecture of CpSADH (Yamamoto et al., 1999; Yang et al., 2014) that has been assigned to two Pfam families (El-Gebali et al., 2019). These are PF08240 (alcohol dehydrogenase GroES-like domain (ADH_N)) and PF00107 (Zinc-binding dehydrogenase (ADH_zinc_N)), respectively. The length distribution among the sequences collected is presented in **Supplementary Figure S1** (see **Supplementary File S1**). Additional 50 sequences were retrieved as above without requiring catalytic evidence but with 3D structure available in the PDB database. We ended up with a set of 280 sequences (there were 41 duplicates among the originally 313 found in the two searches plus the 8 seed sequences) retrieved from the UniProt and PDB databases. The new sequences were added to the seed alignment using MAFFT with Linsi parameters (Katoh and Standley, 2013) and the profile alignment option "--seed" which ensures the curated seed alignment made out of 8 sequences with 3D structures remains preserved. All sequence accessions used are listed in **Supplementary File S2**.

Substrate Binding Pocket Residue-Based Phylogenetic Tree for Alcohol Dehydrogenases

We identified 21 positions with vicinity to substrates within 5 Å in the 3D structure or literature reports for influence on substrate specificity and extracted them from the full length alignment as a 21 site substrate binding pocket profile alignment. This subset was then used to create a phylogenetic tree with MEGA7 (Kumar et al., 2016; Kumar et al., 2018) in order to take into account the relationship of these sequences with sole focus on the potential substrate binding residues. The evolutionary history was inferred by using the Maximum Likelihood method based on the JTT matrix-based model (Jones et al., 1992). Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 2.6014)). The analysis involved 280 amino acid sequences. There was a total of 21 positions in the final dataset.

Statistical Methods to Find Alcohol Dehydrogenase Residues Involved in Substrate Specificity

Further sequence analysis for the examples known to process specific substrates (such as xylitol and L-iditol) included the usage of the multi-Harmony server (Pirovano et al., 2006; Feenstra et al., 2007; Brandt et al., 2010) that uses Sequence Harmony and multi-Relief methods developed to support in the identification of residues with functional specialization within sub-families of proteins as well as Two Sample Logo (Vacic et al., 2006) for additional graphical representation.

RESULTS

Non-Trivial Alignment of Residues in the Vicinity of Substrate in the Sequences of Zinc-Containing Alcohol Dehydrogenase Sequences

Multiple sequence alignments in large protein families across diverse taxonomic kingdoms such as ADHs are technically difficult and often result in misalignments and excessive insertion of gaps. In order to approach this problem, it is crucial to build a reliable sequence alignment that will serve as seed profile alignment for guiding the addition of new sequences belonging to the protein family.

Therefore, few selected, very well studied sequences from *C. parapsilosis* and human origin (Gibbons and Hurley, 2004; Wang et al., 2014; Yamamoto et al., 1999) were analyzed and their

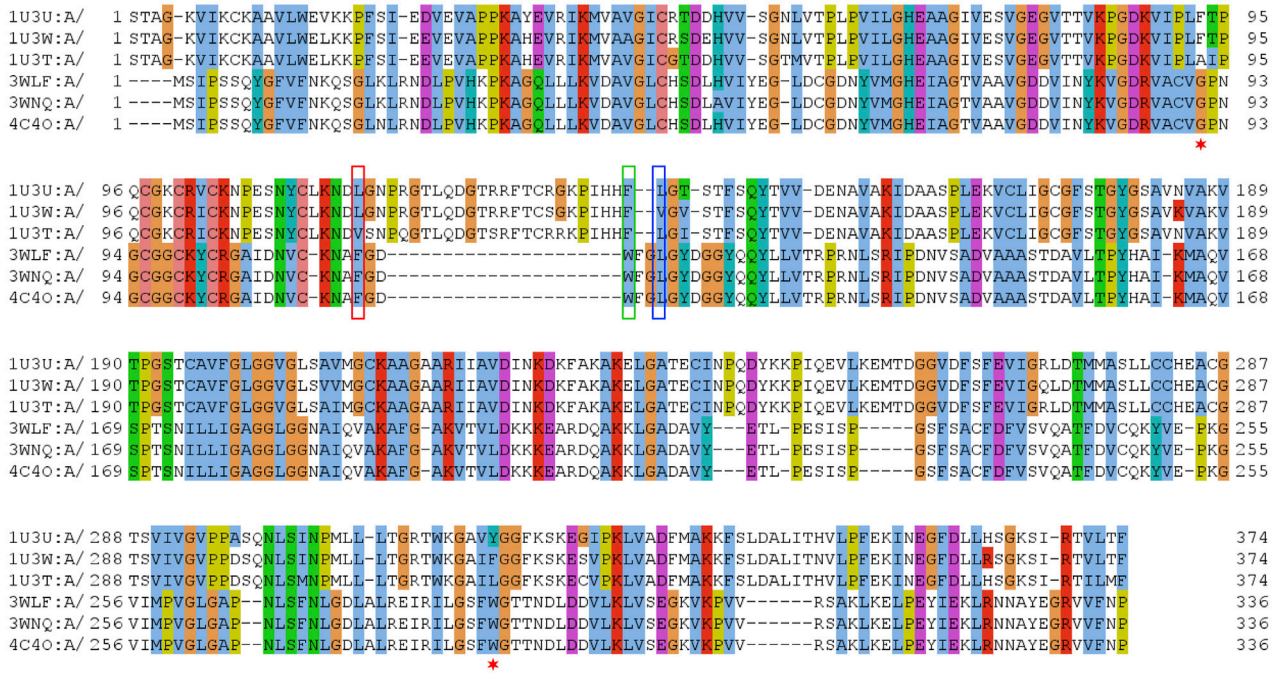

FIGURE 1 | Sequence alignment of selected sequences from zinc-containing ADHs. Manually curated seed alignment with originally eight sequences from PDB entries 1U3U (1U3V has the same sequence within the range shown), 1U3W, 1U3T, 3WLF (3WLE has the same sequence within the range shown), 3WNQ (has an alanine instead of a histidine at position 39 that is part of the set of 21 positions for the substrate binding site) and 4C4O (has an asparagine instead of a lysine at position 19 that is apparently not directly involved in substrate binding). The first sequence set belongs to human, while the second set of four is from *C. parapsilosis*. Three columns are highlighted with red, blue and green boxes. The corresponding sequence positions are part of a loop region, which is structurally different in human and *C. parapsilosis* enzymes. They are positions 12, 14 and 15 in **Figure 2**, respectively. Red asterisks label two residues that were reported in context with substrates' stereo-specificity (Wang et al., 2014).

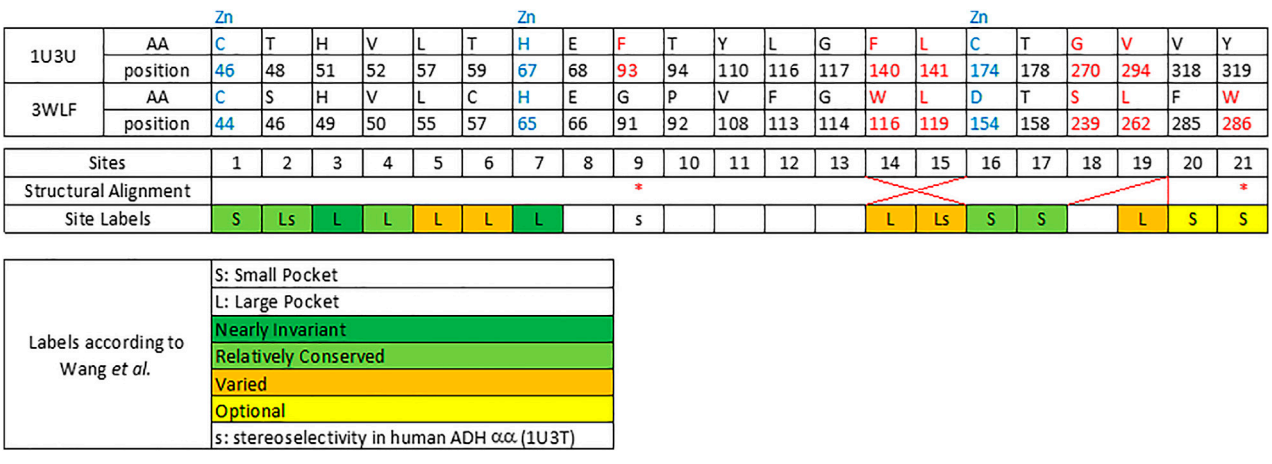

FIGURE 2 | Twenty one sequence positions in close structural proximity to the substrate in zinc-dependent ADH structures. We provide descriptions of 21 sequence positions in the two ADH proteins with structures 1U3U and 3WLF that are in close proximity to the substrate and form the binding site. We refer to annotations of binding site residues from (Wang et al., 2014). Red asterisks label two residues that were reported in context with substrates' stereo-specificity (Wang et al., 2014).

alignment was manually curated based on the available structural information (**Figure 1**). Ultimately, this manually curated and structure-guided alignment is used to identify residues that could play a critical role in substrate specificity of enzymatic reactions. The identification of these positions included the distance criterion to the ligand (having a heavy atom within 5 Å) and literature information (Gibbons and Hurley, 2004; Wang et al., 2014). **Figure 2** summarizes them as sites from 1 to 21.

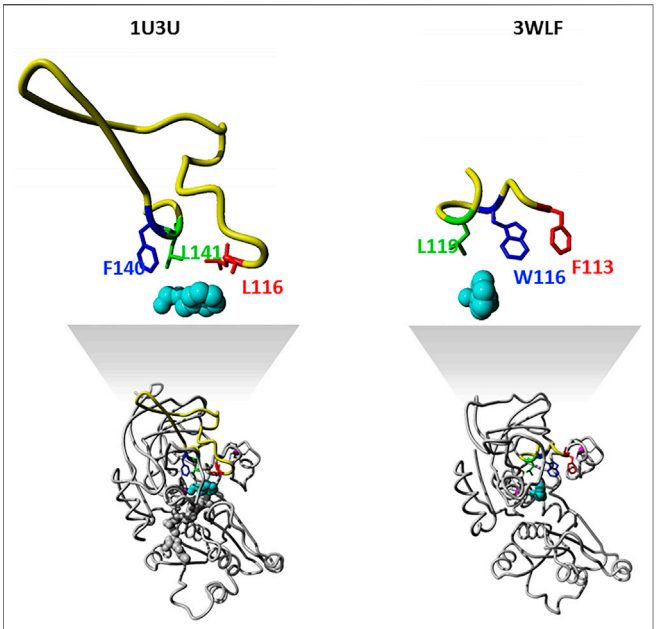

FIGURE 3 | Sequence-structure relationship in the substrate binding pocket of selected zinc-containing ADHs. Structural visualization of protein sequence regions highlighted with boxes in **Figure 1**. The same colour code is followed to indicate the difference observed with swapped positions between structural and sequence alignments. We show both the overall 3D structure (bottom) with the highlighted binding site as well as the loop with the three residues discussed in magnification (upper part).

The zinc-containing ADH family illustrates one of the scenarios where traditional sequence alignment methods are sentenced to fail because the order of certain critical residues in the sequence is not the same in all subgroups. The human ADH sequences (exemplified by 1U3U) have an additional loop between substrate-binding positions (between the red and blue boxes in the alignment shown in **Figure 1**). This loop structure (together with the equivalent segment of the respective *C. parapsilosis* structure 3WLF that does not have this additional loop) is displayed in **Figure 3**. The hinge residues L116, L141 and F140 presented in red, green and blue, are in close proximity to the enzyme substrate. However, these three residues overlap in the structural alignment with F113, W116 and L119 from 3WLF in a way that does not follow a linear sequence as in the primary sequence alignment. Obviously, the loop allows the sequential positions to swap in the structural alignment. Thus, F140 (blue), from the human sequence, structurally overlaps with *C. parapsilosis* L119 (green) instead of W116 (blue).

The core of our approach in this work is to focus on positions that are structurally close to the substrate and to filter them out in a sequence alignment. This insight can then be used to gauge information from other sequences without structural information. The manual curation of the ADH seed alignment allows to include the information that certain columns/positions (displayed in **Figures 1**–**3**) are relevant for subfamily specificity analysis despite being structurally and functionally swapped in some of them. Consequently, this curated seed alignment ensures that these positions are preserved as columns; yet, the respective sub-columns can be swapped in alternative versions of the binding-pocket-only alignment.

Expansion of the Seed Alignment With Further Well-Annotated Sequences From UniProt

Once the seed alignment was created, we added sequences from UniProt already annotated with substrates, catalytic activity and/or 3D structures as described in Methods. We ended up with a final alignment of 280 sequences belonging to the zinc-containing alcohol dehydrogenase family. The final number of different ADH reactions (unique EC numbers) that have been assigned to this family and were simultaneously annotated to have catalytic activity with experimental evidence in UniProt (UniProt Consortium, 2019) was 40 (see EC number list in **Supplementary Table S1** in the **Supplementary Material File S1**). When looking at all possible entries of this family in UniProt (in total, 28,978 examples including both reviewed/un-reviewed and without/with any status of experimental evidence), we found 55 unique reactions listed for them. Hence, only 72.7% of all currently annotated reactions for this family include references to experimental evidence.

From the final alignment, a substrate binding pocket profile was extracted to consider only the 21 positions presented in **Figure 2**. Then, a new substrate binding site tree (**Figure 4**) was created based just on the reduced alignment. The purpose of this tree was to establish the relationship of these sequences focused on the potential substrate specificity. This approach is supported by the fact that enzymes annotated under the same EC numbers and, consequently, support the same reactions are clustered together based on these 21 residues. With this tool in hands, we can attempt to classify unannotated ADHs based on their binding pocket properties.

Classification of Alcohol Dehydrogenase Sequences With Xylitol and L-Iditol as Substrates Based on Their Binding Pocket Properties

To exemplify the utility of clustering sequences based on their substrate binding pocket, we highlight the subtree branches with the enzymes D-xylulose reductase with the substrate xylitol and L-iditol 2-dehydrogenase with substrates L-iditol and xylitol (**Figures 5A,B**). These substrates are both linear sugar alcohols with xylitol being shorter by one carbon and hydroxyl group.

To identify residues that could be linked to the substrate specificity, we took the 21 binding pocket alignment of the sequences belonging to these two enzyme families in the subtree and compared the two groups with Sequence Harmony. This indicated a position with a preference for methionine (M) in position 19 for the Xylitol family and leucine (L) for the L-iditol family. Notably, both residues are quite similar in side chain volume (166.7 $Å^3$ versus 162.9 $Å^3$ respectively; (Zamyatnin, 1972; Zamyatnin, 1984)); yet, leucine is branched and methionine has the longer side chain. Showing the position in an example structure of the enzyme in complex

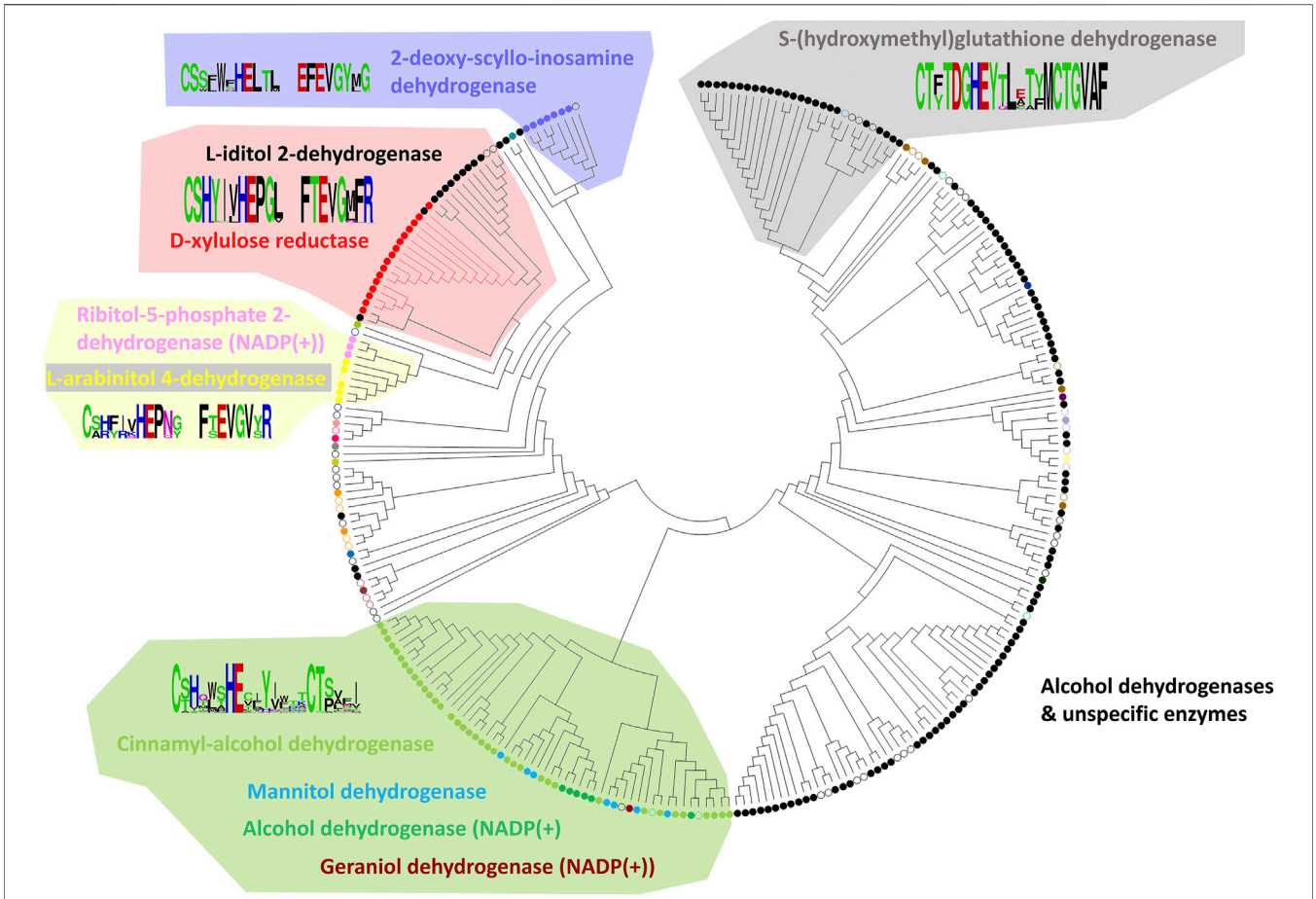

FIGURE 4 | Substrate binding site tree based on substrate binding residues of ADHs with annotated EC and experimental evidence for catalytic activity. What looks like a phylogenetic tree in this figure is actually a substrate binding site tree created with phylogenetic tree tools from the abbreviated alignment containing only designated residues from the substrate binding pocket. The tree with the highest log likelihood (−3,973.88) is shown. Unique and identical EC numbers are labelled with same colored dots next to the branch. Any enzyme annotated either as 1.1.1.1, 1.1.1.- or had more than one EC number associated with it was colored black and considered promiscuous/unspecific. Enzyme groups with EC numbers having specific substrate annotation are highlighted with the respective dot color and labelled with the substrate name. For EC number annotation, see **Supplementary Table S1**. The specific sub-branch used in the example for **Figure 5** is colored red. For several branches of the tree, we show the sequence logo (Vacic et al., 2006) for the binding site residues within the relevant sequence group. A linear, high-resolution version of this tree with explicit EC number annotation of all branches is available as **Supplementary Figure S3** in **Supplementary Files S1, S3**. For comparison, we also provide the true phylogenetic tree of ADHs generated from the full-length alignment (see **Supplementary Figure S4** in **Supplementary Files S1, S4**).

with co-factor and ligand, one can see that the longer M residue results in a tighter binding pocket and L leaving more space (**Figure 5C**). This is consistent with M being the preferred residue for the shorter ligand xylitol and L for the longer L-iditol. Thus, the substrate binding pocket of L-iditol 2-dehydrogenase with L provides more space and can more easily accept both long and short substrates.

The binding pocket subtree from these enzymes (**Figure 5B**) shows that there is not a clean split into phylogenetic groups matching the enzyme codes but the M/L preference is statistically significant (Z-score −4.02). Within the subtree, there is a further division into the big group with M/L and a smaller subgroup that has mostly M at position 19. For comparison, if we run the Sequence Harmony analysis over the full sequence, we get a long list of 187 candidate positions ($Z < -3$) while the focus on the 21 binding pocket residues made it easy to spot the two main candidates.

Identifying Relevant Binding Pocket Positions for Mutations in (R,R)-Butanediol Dehydrogenase From *Bacillus Subtilis*

Another example of identifying relevant positions based on sequences found in the binding pocket that participate in ligand binding is illustrated here for the enzyme (R,R)-butanediol dehydrogenase from *Bacillus subtilis* (UniProt O34788/BDHA_BACSU).

In the phylogenetic tree (**Figure 4**), this enzyme family is neighbored by two other branches containing L-arabinitol 4-dehydrogenases and ribitol 5-phosphate 2 dehydrogenases on one hand and 2-deoxy-scyllo-inosamine dehydrogenases on the other. A new sequence alignment was generated to account for this set only. The sequence analysis suite Sequence Harmony (Pirovano et al., 2006; Feenstra et al., 2007; Brandt et al., 2010)

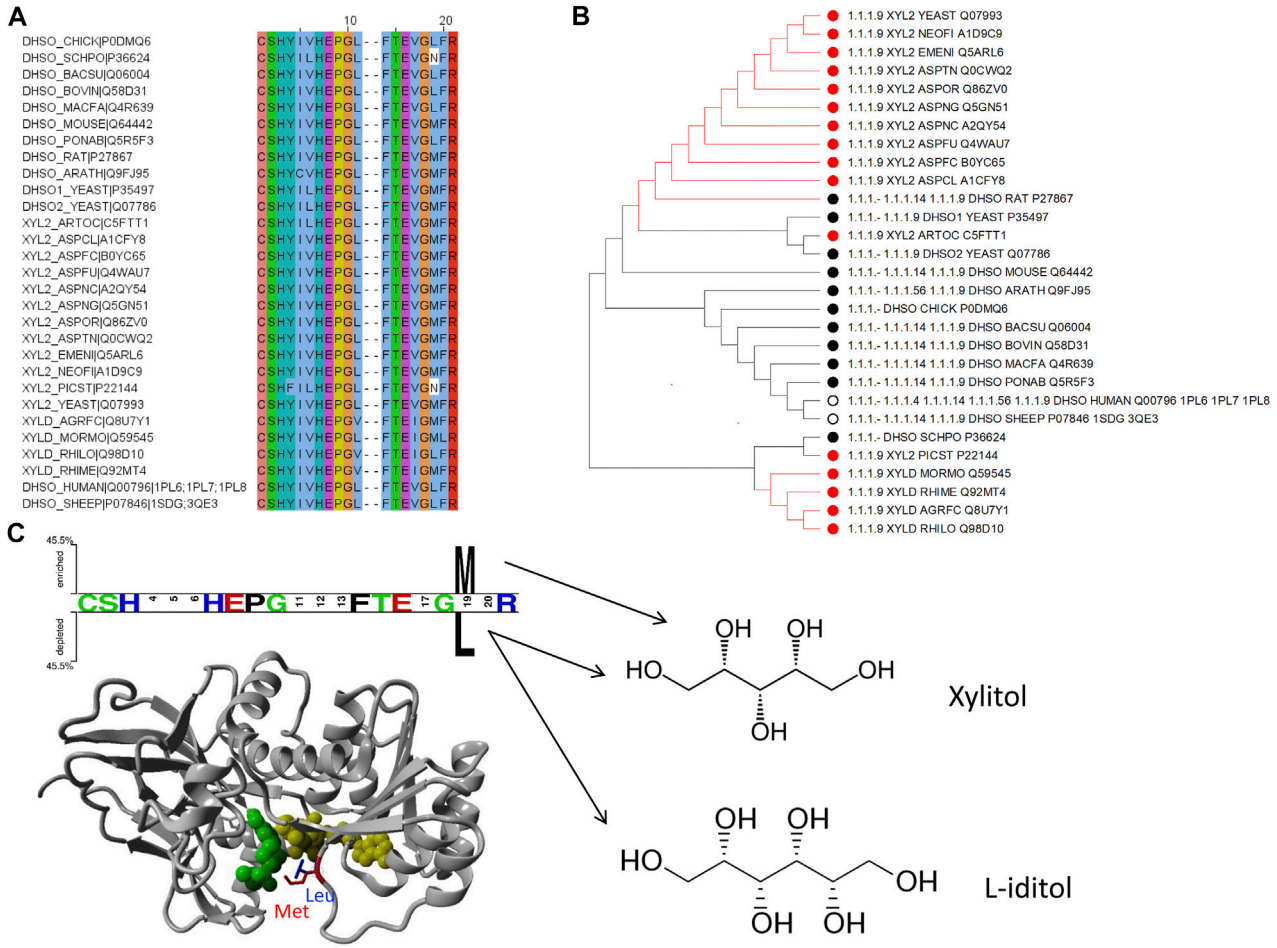

FIGURE 5 | Example of how to identify residues with functional specialization within sub-families of proteins: the D-xylulose reductase (EC 1.1.1.9) and L-iditol 2-dehydrogenase (EC 1.1.1.14) cases. **(A)** Sequence alignment with positions from **Figure 2** to include enzymes from EC 1.1.1.9 and 1.1.1.14 with experimental verification as annotated in UniProt only. **(B)** Subtree from **Figure 4** to consider branch with the two enzymes in question. The color coding of dots next to the branch corresponds to the EC numbers. **(C)** Two Sample Logo (Vacic et al., 2006) and Structural representation of PDB ID 1PL6 with a modelled methionine in position 19 (from **Figure 2**) for visualization purpose.

was used to identify which positions would represent a significant difference between the three highlighted branches. **Figure 6A** summarizes the sequence positions for the three protein groups in the branch for (R,R)-butane-2,3-diol dehydrogenase. Note that the protein P39713/BDH2_YEAST in the same branch has been annotated to catalyze the reaction (R)-acetoin + NAD (+) ≤> diacetyl + NADH. **Figure 6B** shows the results of Sequence Harmony, a tool that supports identification of the residues that are different among the potential binding positions of the three branches aligned.

Identification of Putative Thermophile Organisms With Similar CpSADH

In biotechnology, usage of enzymes from extremophiles is often preferred as they can execute their function despite of a harsher handling. We note that the approach presented here can also be useful to identify thermophile organisms that potentially harbor enzymes catalyzing the reaction of interest. This opportunity is exemplified here by using the CpSADH as reference if one is interested in identifying possible thermostable alcohol dehydrogenases (ADHs).

524 protein sequences from UniProt identified with the domain architecture (with sequence domains PF08240 or PF00107) and belonging to an organism identified as thermophile (using the list of thermophile organisms provided by Dr. Igor Berezovsky (Zeldovich et al., 2007; Ma et al., 2010)) were merged with the set of reviewed protein entries that were retrieved from UniProtKB under the family annotation "zinc containing alcohol dehydrogenase family", EC:1.1.1.* with catalytic activity and experimental evidence annotated. These, together with a few reference PDB sequences, generated a list of 808 protein sequences. These were aligned with the MAFFT algorithm (using Linsi parameters) (Katoh and Standley, 2013) for a full-length sequence alignment (see below section "A"). For generating the alignment of only the residues belonging to or

A

binding pocket positions		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
3WLF	AA	C	S	H	V	L	C	H	E	G	P	V	F	G	W	L	D	T	S	L	F	W
	position	44	46	49	50	55	57	65	66	91	92	108	113	114	116	119	154	158	239	262	285	286
BDHA_BACSU	AA	C	S	H	E	T	E	H	E	P	I	L	G	-	F	L	E	V	G	I	I	Y
	position	37	39	42	43	60	62	70	71	96	97	109	114		115	118	152	156	245	268	291	293
BDH1_YEAST	AA	C	S	H	E	S	A	H	E	A	A	L	G	-	F	L	E	V	G	V	I	Y
	position	39	41	44	45	63	65	73	74	99	100	130	135		136	139	173	177	269	292	316	318
BDH2_YEAST	AA	C	T	H	E	S	N	H	E	P	T	I	G	-	L	A	Q	V	G	M	M	Y
	position	39	41	44	45	63	65	73	74	99	100	130	135		136	139	173	177	270	293	317	319

B

BDHA_BACSU's positions	37	39	42	43	60	62	70	71	96	97	109	114	-	115	118	152	156	245	268	291	293
sequence harmony (Z-score)	nan	-2.73	-3.89	-4.66	-6.07	-3.24	nan	nan	-3.53	-4.49	-3.83	-2.28	-0.01	-3.86	-5.22	-0.8	-0.02	nan	-3.81	-4.51	-3.11
sequence harmony (score)	1	0.764	0.449	0.27	0.067	0.134	1	1	0.579	0.291	0.274	0.341	0.876	0.449	0.101	0.915	0.9	1	0.397	0.163	0.449
suggested sites for mutation				*	*	*				*					*					*	

C

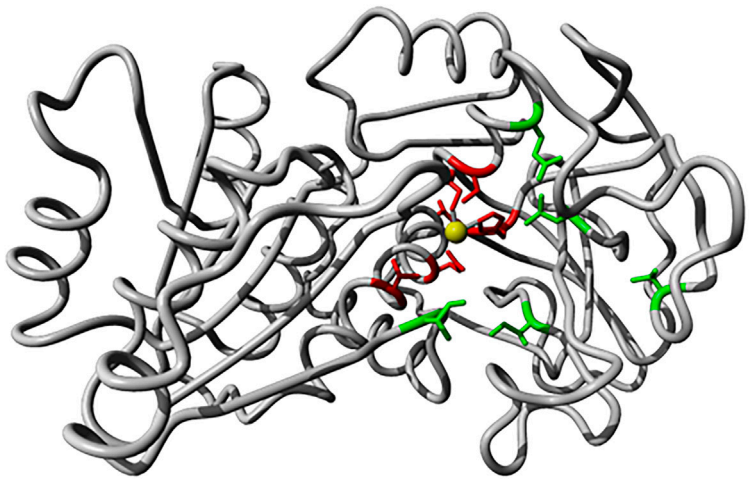

FIGURE 6 | Binding pocket positions for mutations in (R,R)-butanediol dehydrogenase from *Bacillus subtilis*. **(A)** Alignment of 21 sequence positions as in **Figure 2** with (R,R)-butanediol dehydrogenases (BDHA_BACSU, BDH1_YEAST) and the additional probable diacetyl reductase [(R)-acetoin forming] 2 (BDH2_YEAST) having the structure 3WLF as reference. **(B)** Results of Sequence Harmony scores and sites for suggested mutations highlighted with asterisk. Sequence Harmony scores were coloured from green to red according to a gradient from smaller to bigger numbers, respectively; where red represents no significant difference observed for the positions of the enzymes in the neighbouring branches, while green highlights the positions of interest. These are the most probable sites relevant for the enzyme substrate specificity. **(C)** X-ray crystal structure 6IE0 (Wang et al., 2018) of the (2R, 3R)-butanediol dehydrogenase where the positions with a significant Sequence Harmony score are highlighted in green (43, 60, 97, 118 and 291) and those with basically no significance are highlighted in red (37, 39, 70, 71, 152, 156). The zinc ion is shown in yellow. Note that the red sites are the ones localized to the catalytic site and are conserved throughout, while the green ones are the ones likely interacting with other parts of the substrate relevant for its specificity.

being close to the binding pocket, we used the MAFFT–add option to add the unaligned sequences to our manually curated alignment taking into account the structural position of some relevant binding pocket residues (see below section "B").

The reason for exploring these two alignment options is that the curated alignment forces any of the other sequences to fit to it. Given the complexity of loops and structural rearrangement, the overall sequence alignment might be compromised in certain situations. The idea of combining all these methods is to ultimately come up with a few thermophile species that could potentially be interesting to be explored further, taking into account how they relate both at full sequence and binding pocket levels.

A) Full sequence alignment option: The evolutionary analyses were conducted in MEGA7 (Kumar et al., 2016). In addition to two schemes of handling alignment gaps and missing/ambiguous data (either excluding all sites with gaps or sites with less than 95% coverage), two alternative methods were used for tree construction–the neighbor-joining method (Saitou and Nei, 1987) (see **Supplementary Figures S2A,B**) and Maximum Likelihood method based on the JTT matrix-based model (Jones et al., 1992) (see **Supplementary Figures S2C,D**).

In all four scenarios (see trees in **Supplementary Figures S2A–D**), the closest protein to CpSADH is ACM07214/

TABLE 1 | ADHs in thermophile organisms closest to CpSADH.

Organism	Protein accession code
Thermomicrobium roseum	WP_012643201.1
Thermomonospora curvata	WP_012853139.1
Thermobifida fusca	WP_011293194.1
Thermobifida fusca	WP_011291920.1
Thermobifida fusca	WP_016188671.1
Moorella thermoacetica[a]	WP_069588064.1
Alicyclobacillus acidocaldarius[a]	WP_012811644.1

[a]*Organism not displayed in all versions of tree branches.*

WP_012643201 from *Thermomicrobium roseum*. When alignment columns with gaps in some sequences were included (this increases the number of positions with phylogenetic signal), proteins from *Thermobifida fusca* and *Thermomonospora curvata* consistently appear in the same branch regardless of the tree-building method. With further relaxation (after removing all alignment columns with gaps occurring), additional sequences are hit in the search and proteins from *Alicyclobacillus acidocaldarius* and, in one case, *Moorella thermoacetica* can be found.

B) Alignment restricted to binding pocket residues: When using the procedure described in the Methods section to align new sequences to the manually curated sequences, we still retrieve the species *T. roseum*, *T. fusca* and *T. curvata* in the evolutionary analysis depending on the tree-building method used (**Supplementary Figures S2E,F**). Thus, the ADHs from these three thermophile organisms are the closest to CpSADH. They (as listed in **Table 1**) are recommended to be further experimentally explored.

Long Chain Alcohol Oxidases

The workflow applied above to ADHs and some of their subgroups is a quite general methodology and provides a framework to classify candidate enzymes by their substrate preferences as well as highlights the position and type of residues directing substrate specificity. To show its

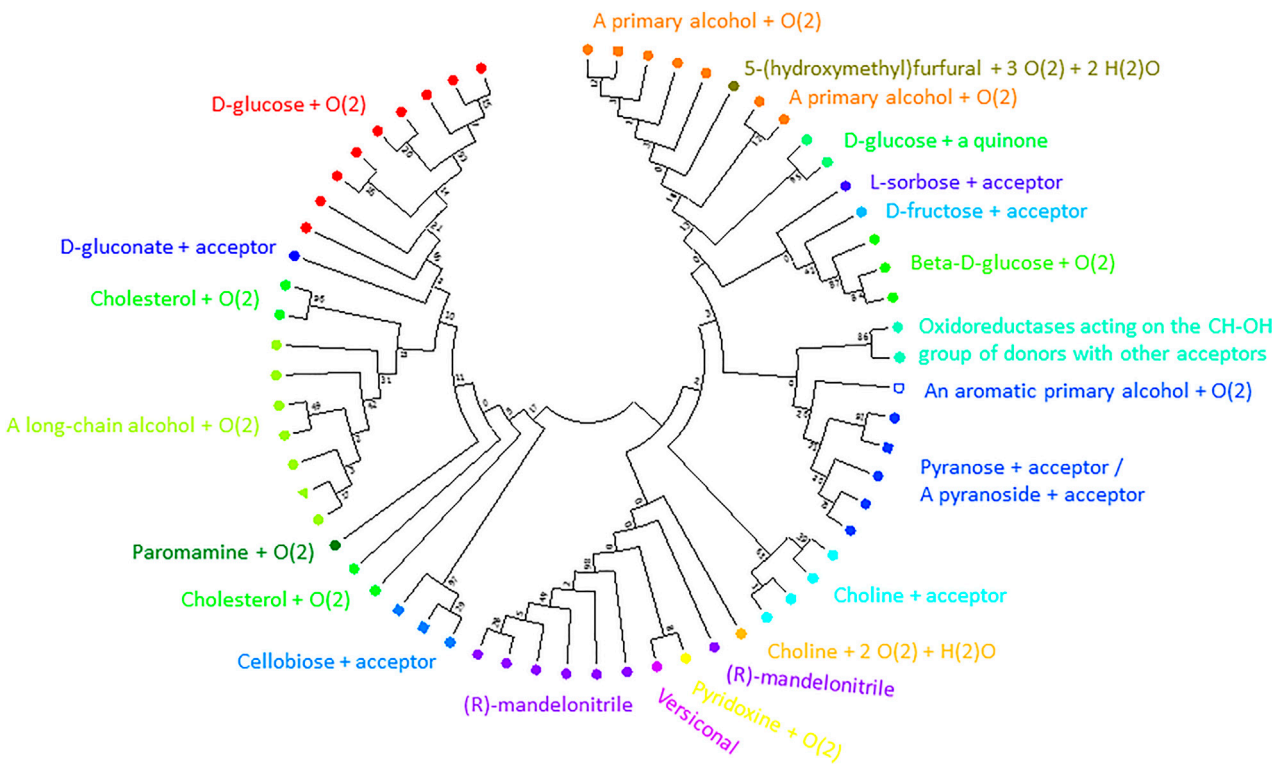

FIGURE 7 | Substrate binding site tree based on substrate binding residues of AOxs with annotated EC and experimental evidence for catalytic activity. What looks like a phylogenetic tree in this figure is actually a substrate binding site tree created with phylogenetic tree tools from the abbreviated alignment containing only designated residues from the substrate binding pocket. The tree's branches are grouped according to the binding pocket description based on positions identified to be within 5 Å to the inhibitor (ABL) of 1NAA (Hallberg et al., 2003). The same methodology described for the ADHs was applied using MEGA (Kumar et al., 2016). A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 5.7490)). This analysis involved 62 amino acid sequences. There were a total of 19 sequence positions (supposedly involved in substrate binding) in the final dataset. The color code corresponds to EC numbers. The green triangle marks the query sequence. For EC number annotation, see **Supplementary Table S1**. We show the true phylogenetic tree derived from the full-length alignmnents of AOxs as **Supplementary Figure S5** in **Supplementary Files S1, S5**. **Supplementary Figure S6** (in **Supplementary Files S1, S6**) is a high-resolution version of **Figure 7** with EC-number annotated branches.

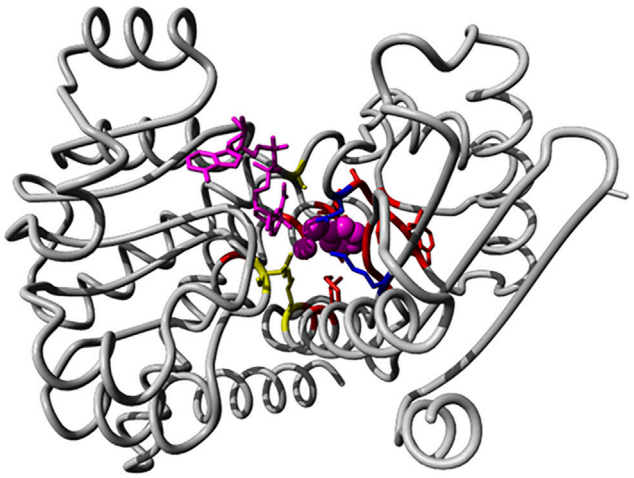

FIGURE 8 | Binding pocket positions in the AmDH reference structure 1C1D. We show the 3D structure view of AmDH reference (chain A of the PDB entry 1C1D) with ligands in magenta (substrate as sphere and co-factor as stick representations). Side chain of residues within 5 Å of the binding pocket are displayed in stick mode. Residues described as catalytic are shown in blue (K78 and D118 (Vanhooke et al., 1999)), while positions (K66, S149 and N262) described to play a role in enantioselectivity (Ye et al., 2015) are shown in yellow.

straightforward applicability, we show its application for the enzyme family of long chain alcohol oxidases.

We started the reference sequence from *Candida tropicalis* under the UniProt accession code Q9P8D9 (Vanhanen et al., 2000) with the domain architecture assigned to two Pfam families: PF00732 (GMC_oxred_N) and PF05199 (GMC_oxred_C), belonging to the glucose-methanol-choline oxidoreductase family (GMC oxidoreductase).

Our UniProt search requiring proteins to belong to the GMC oxidoreductase family and to be annotated to have catalytic activity with experimental evidence retrieved additional 56 sequences. Because the reference sequence did not have a PDB structure available, sequence search with tools such as BLASTP (Altschul et al., 1997; Johnson et al., 2008) and HHpred (Zimmermann et al., 2018) were used to identify similar structures with the coordinates of a possible substrate resolved. We found five suitable examples: 1KDG, 1NAA (Hallberg et al., 2003), 4H7U (Tan et al., 2013), 5HSA (Koch et al., 2016), and 5OCI (Carro et al., 2017). We added sequences from all five structures to the seed sequence alignment (our final dataset increased to 61 proteins). All sequence accessions are listed in **Supplementary File S2**.

1NAA, a cellobiose dehydrogenase flavoprotein, has the coordinates of its inhibitor resolved (Hallberg et al., 2003). Its active site is structurally similar to that of glucose and cholesterol oxidases whose mechanism of oxidation is still poorly understood (Hallberg et al., 2003). We identified 19 positions sites within 5 Å of its inhibitor ABL (278, 279, 282, 297, 310, 312, 562, 563, 584, 586, 590, 607, 609, 686, 687, 688, 689, 732, 733; numbering in accordance with 1NAA sequence) potentially involved in interaction with the enzyme's substrate. At this point, we cannot say whether we identified all residues relevant for substrate binding given the limited structural information. Finally, we constructed a binding site tree (**Figure 7**) derived from just from binding pocket positions (see the true phylogenetic tree for comparison in **Supplementary Figure S5** in **Supplementary Files S1, S5**). Satisfactorily, sequences with same EC numbers cluster, as a trend, in the same branch and, thus, co-clustering uncharacterized sequences can be assumed have the same or similar substrate specificity.

Amino Dehydrogenases

We identified 66 members of the (Glu, Leu, Phe, Val)-dehydrogenases family with catalytic activity annotated by experimental evidence and with sequence lengths between 200 and 500 from the UniProt database. All sequence accessions are listed in **Supplementary File S2**. The exhaustive family search for AmDH sequences in the NR database at NCBI using HMMER identified almost 26,000 candidates. We classified the enzymes with known activity in a phylogenetic tree based on full length sequences and colored and annotated them by their Enzyme Classification (EC) code (see **Supplementary Figure S7**). Then, we analyzed the reference 3D structure 1C1D (Brunhuber et al., 2000) from the Protein Data Bank RSCB (Goodsell et al., 2020) and literature sources (Ye et al., 2015). We then identified 20 positions (38, 39, 40, 63, 66, 67, 78, 114, 115, 116, 117, 118, 137, 149, 262, 288, 291, 292, 295, and 296 in chain A of 1CID; see **Figure 8**) relevant for substrate binding and specificity.

After extracting the sub-alignment with only those positions, we constructed a new phylogenetic tree. In this classification based on the binding pocket residues, the enzymes with known activity cluster, as trend, with those having similar substrate specificity (see **Figure 9**). We find that the different dehydrogenases are much better grouped in branches of the phylogenetic tree with regard to substrate specificity when just the binding pocket positions are considered (and not the full sequences) and this can be especially clearly seen for the phenylalanine/valine dehydrogenases (as two branches for the tree from the full-length alignment with the valine dehydrogenase branch in the middle but as one branch in the tree from the binding pocket sequence alignment).

DISCUSSION

It was not our goal to exhaustively classify the large enzyme families (ADHs, AOXs, AmDHs) along substrate specificity but rather to show how groups of sequences potentially useful for subsequent enzyme engineering could be selected with quite limited effort.

To facilitate the classification of large bodies of enzyme sequences with regard to their potential substrates, we suggest a tree-step procedure in this work. First, it is necessary to understand what the residues that interact with the substrate are. This information can be gathered from the scientific

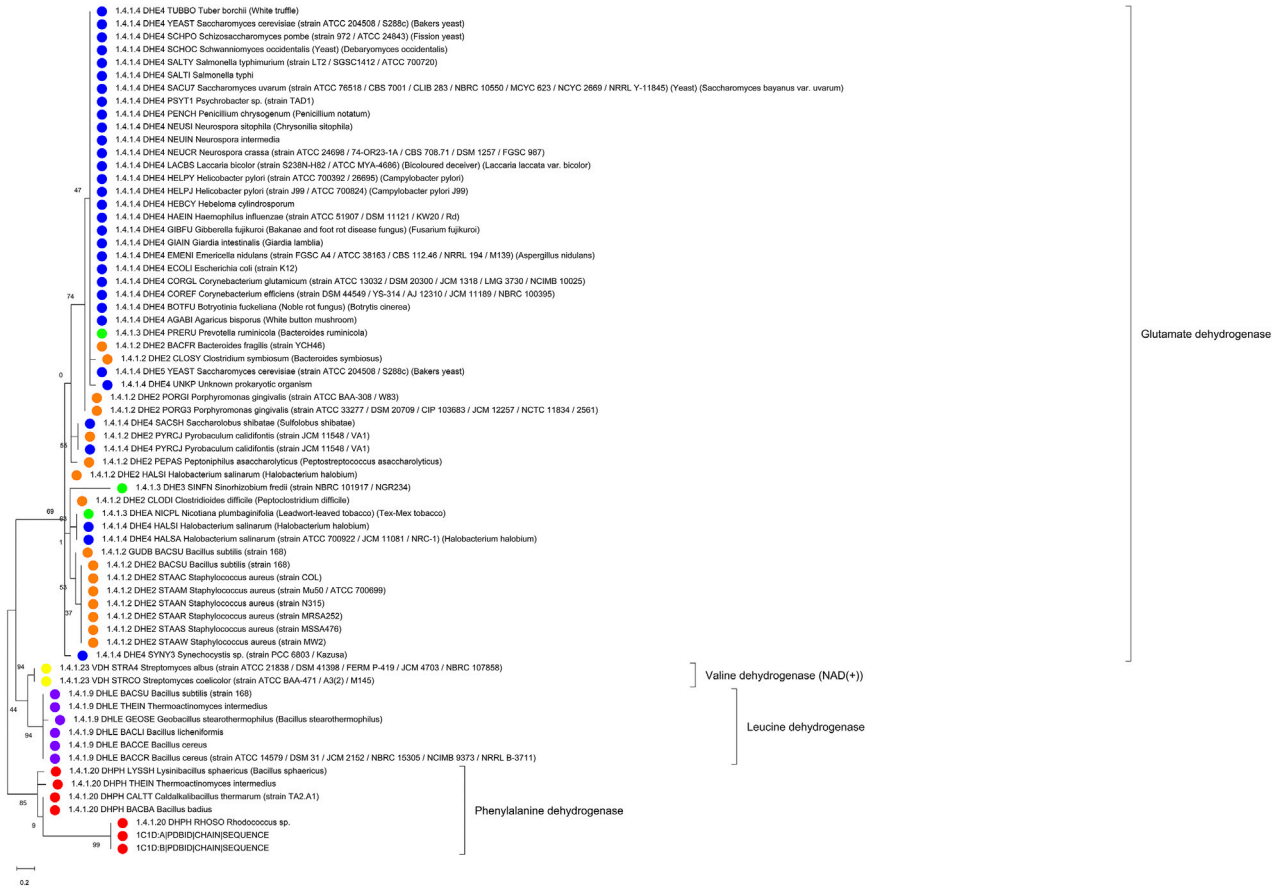

FIGURE 9 | Substrate binding site tree based on substrate binding residues of AmDHs with annotated EC and experimental evidence for catalytic activity. What looks like a phylogenetic tree in this figure is actually a substrate binding site tree created with phylogenetic tree tools from the abbreviated alignment containing only designated residues from the substrate binding pocket. Evolutionary analyses were conducted in MEGA (Kumar et al., 2018). We present the results of molecular phylogenetic analysis of the AmDH binding pocket by the Maximum Likelihood method. The evolutionary history was inferred by using the Maximum Likelihood method and JTT matrix-based model (Jones et al., 1992). The tree with the highest log likelihood (-295.80) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the JTT model, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 2.0446)). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. This analysis involved 66 amino acid sequences. There were a total of 20 positions in the final dataset. The color code corresponds to EC numbers (yellow–valine dehydrogenases, red–phenylalanine dehydrogenases, violet–leucine dehydrogenases, all others–glutamate dehydrogenases). For comparison, see **Supplementary Figure S7** with the tree derived for the full-length sequence alignment (in **Supplementary File S1**).

literature about the enzyme's family, from annotations in sequence databases or, most directly, from 3D structures of the respective enzymes together with substrates, cofactors, etc. if available. Clearly, diverse substrates processed by the same family of enzymes (for example, larger or smaller ones) will have different binding residue list. Surely, some binding site positions might be more significant for specificity than others. We recommend to include all residues that have a role at least for some substrate types. Very often in practical applications, the available information is incomplete in this regard and, as a trend, the list of residues put together in this effort will be incomplete. The list of binding site residue positions is also the first recommendation for mutations to test in enzyme engineering efforts.

In a second step, a group of sequences is gathered from the protein sequence and literature databases that is annotated with substrate specificity data. Together with the sequences from the first step, the joint sequence set (after removal of potential duplications) is used for the creation of a seed multiple sequence alignment that emphasizes residues from various sequences with equal role in substrate binding being put into the same alignment column. At some stage, this process will require manual interference beyond the automatism of alignment programs, for example if the sequential order and 3D structural role are conflicting as in the case of our ADH seed alignment.

The third step involves the creation of a "binding site tree" or binding cavity similarity tree" when just the list of binding site residues is used as input for phylogenetic tree-generating

tools. It is expected that sequences with similar substrate specificity will be grouped together in the same branch of the tree. Any amount of non-annotated sequences can now be mapped onto the tree with sequence similarity criteria (for example by just expanding the multiple sequence alignment used for tree construction with the constraint of preserving the seed alignment) and, again, we presume that the new sequences associated with a given branch will have the same or very similar substrate specificity as the annotated sequences located there.

Clearly, the procedure is only for predicting and hypothesizing about substrate specificity of non-annotated sequences. The method is not a panacea. If amino acid residue patterns show clear trends within and between branches (the patterns reflect the physics of the respective binding style), the predictions will be more likely true. There are many caveats: On the one hand, the information about possible substrates in the literature can be incomplete, even erroneous and certain applicable compounds might be missing in the tree's annotation. On the other hand, the sequences can represent promiscuous enzymes or even enzymatically non-active binders. The ADH family as described in the introduction is a pertinent illustration for all these possibilities.

Since the "binding site" tree is constructed from only a minor fraction (maybe, two dozens of positions) of the full-length multiple sequence alignment of the enzyme family (that, typically, encompasses a few hundred positions), there should be no surprise if the binding site tree and the true phylogenetic tree (generated from the full-length alignment) do not share much similarity. Additionally, bootstrap values and other statistical measures in the output for the "binding site" tree might have diminished significance because of the smallness of the number of involved alignment positions. All this can be easily disregarded for the purpose of hypothesis generation on substrate specificity for uncharacterized sequences.

Yet, the reader will agree that, intriguingly, the "binding site" tree and the true phylogenetic tree are surprisingly similar (see **Figure 4** and **Supplementary Figure S4** for ADHs, **Figure 7** and **Supplementary Figure S5** for AOXs, **Figure 9** and **Supplementary Figure S7** for the AmDHs). That means that the few positions involved in substrate binding and catalysis usually contain most of the information that is otherwise contained in the full-length multiple sequence alignment. Further, the appearance of the same substrate specificity in disconnected branches of the phylogenetic tree can and does happen (as it takes just a few mutations compared with a large number of mutations necessary for changes in the overall tertiary structural arrangements and fitting) but, nevertheless, it is not a very frequent event. The quantification of this evolutionary phenomenon would be an interesting scientific task on its own. At the same time, analyses of partial sequence alignments from selected branches of the phylogenetic tree with different and equal substrate specificity, with conflicting or re-occurring EC numbers can give hints about additional residue positions that might have a role in determining the suitable substrate compounds (and, thus, improve the "binding site" tree).

CONCLUSION

Zinc-dependent alcohol dehydrogenases (ADHs), long chain alcohol oxidases (AOxs) and amino dehydrogenases (AmDHs) are large enzyme families with good potential for biocatalysis applications through directed evolution towards new substrates and reactions. By creating a tree based on the alignment of potentially substrate binding sequence positions using a combination of bioinformatics tools, we can systematically classify these sequences (both characterized and uncharacterized ones) relative to known substrates. As a trend, enzymes with similar substrates will reside in the same branch of the tree. The sequence subgroup identified in this manner becomes more manageable and the most suitable, naturally occurring enzyme and the most relevant sites for substrate specificity can then be targeted for the intended engineering purpose.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: The datasets are publicly available and can be retrieved from sequence and structure databases (Genebank/EMBL, PDB). Requests to access these datasets should be directed to FS, fernanda@bii.a-star.edu.sg.

AUTHOR CONTRIBUTIONS

ZL confronted the other partners in the project with the problem of the global analysis and substrate-specific sub-classification of the ADH, AOx and AmDH sequence families. FS and SM-S with inputs from BE and FE conceived the methodical approach. All computations were done by FS. All authors analyzed and evaluated the results from this work. The manuscript was written by FS, FE and BE. All authors read and approved the article.

FUNDING

This research was supported by the National Research Foundation (NRF), Singapore, through a Competitive Research Programme (CRP) (Project ID NRF-CRP17-2017-03) to ZL and BE.

ACKNOWLEDGMENTS

The authors acknowledge general support from A*STAR.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2021.701120/full#supplementary-material

A Supplementary File S1 (aka "Data Sheet 1.PDF") with supplementary material including Supplementary Figures (S1, S2A-F, S3-S7) and Table S1 is available for this article. Supplementary File S2 (aka "Table 1.XLSX") contains the accession numbers of all sequence sets described in this article in electronically readable format. Supplementary Figures S3-S6 can also be accessed as high-resolution image in Supplementary Files S3-S6 (aka "Data Sheet 2-5.PDF"). Additionally, a Supplementary File S7 "alignments" (a zipped directory, aka "Data Sheet 6.ZIP") with fasta-formatted alignments for the binding site positions of ADHs, AOxs and AmDHs is provided.

REFERENCES

Altschul, S., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a New Generation of Protein Database Search Programs. *Nucleic Acids Res.* 25, 3389–3402. doi:10.1093/nar/25.17.3389

Arnold, F. H. (2019). Innovation by Evolution: Bringing New Chemistry to Life (Nobel Lecture). *Angew. Chem. Int. Ed.* 58, 14420–14426. doi:10.1002/anie.201907729

Brandt, B. W., Feenstra, K. A., and Heringa, J. (2010). Multi-Harmony: Detecting Functional Specificity from Sequence Alignment. *Nucleic Acids Res.* 38, W35–W40. doi:10.1093/nar/gkq415

Brunhuber, N. M. W., Thoden, J. B., Blanchard, J. S., and Vanhooke, J. L. (2000). Rhodococcusl-Phenylalanine Dehydrogenase: Kinetics, Mechanism, and Structural Basis for Catalytic Specifity†,‡. *Biochemistry* 39, 9174–9187. doi:10.1021/bi000494c

Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G. V., et al. (2021). RCSB Protein Data Bank: Powerful New Tools for Exploring 3D Structures of Biological Macromolecules for Basic and Applied Research and Education in Fundamental Biology, Biomedicine, Biotechnology, Bioengineering and Energy Sciences. *Nucleic Acids Res.* 49, D437–D451. doi:10.1093/nar/gkaa1038

Carro, J., Martínez-Júlvez, M., Medina, M., Martínez, A. T., and Ferreira, P. (2017). Protein Dynamics Promote Hydride Tunnelling in Substrate Oxidation by Aryl-Alcohol Oxidase. *Phys. Chem. Chem. Phys.* 19, 28666–28675. doi:10.1039/c7cp05904c

de Smidt, O., du Preez, J. C., and Albertyn, J. (2008). The Alcohol Dehydrogenases ofSaccharomyces Cerevisiae: a Comprehensive Review. *FEMS Yeast Res.* 8, 967–978. doi:10.1111/j.1567-1364.2008.00387.x

Dunn, P. J. (2012). The Importance of green Chemistry in Process Research and Development. *Chem. Soc. Rev.* 41, 1452–1461. doi:10.1039/c1cs15041c

Eisenhaber, B., Sinha, S., Wong, W.-C., and Eisenhaber, F. (2018). Function of a Membrane-Embedded Domain Evolutionarily Multiplied in the GPI Lipid Anchor Pathway Proteins PIG-B, PIG-M, PIG-U, PIG-W, PIG-V, and PIG-Z. *Cell Cycle* 17, 874–880. doi:10.1080/15384101.2018.1456294

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., et al. (2019). The Pfam Protein Families Database in 2019. *Nucleic Acids Res.* 47, D427–D432. doi:10.1093/nar/gky995

Feenstra, K. A., Pirovano, W., Krab, K., and Heringa, J. (2007). Sequence harmony: Detecting Functional Specificity from Alignments. *Nucleic Acids Res.* 35, W495–W498. doi:10.1093/nar/gkm406

Gibbons, B. J., and Hurley, T. D. (2004). Structure of Three Class I Human Alcohol Dehydrogenases Complexed with Isoenzyme Specific Formamide Inhibitors†,‡. *Biochemistry* 43, 12555–12562. doi:10.1021/bi0489107

Goodsell, D. S., Zardecki, C., Di Costanzo, L., Duarte, J. M., Hudson, B. P., Persikova, I., et al. (2020). RCSB Protein Data Bank: Enabling Biomedical Research and Drug Discovery. *Protein Sci.* 29, 52–65. doi:10.1002/pro.3730

Goswami, P., Chinnadayyala, S. S. R., Chakraborty, M., Kumar, A. K., and Kakoti, A. (2013). An Overview on Alcohol Oxidases and Their Potential Applications. *Appl. Microbiol. Biotechnol.* 97, 4259–4275. doi:10.1007/s00253-013-4842-9

Hallberg, B. M., Henriksson, G., Pettersson, G., Vasella, A., and Divne, C. (2003). Mechanism of the Reductive Half-Reaction in Cellobiose Dehydrogenase. *J. Biol. Chem.* 278, 7160–7166. doi:10.1074/jbc.m210961200

Henehan, G. T. M., and Oppenheimer, N. J. (1993). Horse Liver Alcohol Dehydrogenase-Catalyzed Oxidation of Aldehydes: Dismutation Precedes Net Production of Reduced Nicotinamide Adenine Dinucleotide. *Biochemistry* 32, 735–738. doi:10.1021/bi00054a001

Höög, J.-O., Strömberg, P., Hedberg, J. J., and Griffiths, W. J. (2003). The Mammalian Alcohol Dehydrogenases Interact in Several Metabolic Pathways. *Chemico-Biological Interactions* 143-144, 175–181. doi:10.1016/s0009-2797(02)00225-9

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., and Madden, T. L. (2008). NCBI BLAST: a Better Web Interface. *Nucleic Acids Res.* 36, W5–W9. doi:10.1093/nar/gkn201

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The Rapid Generation of Mutation Data Matrices from Protein Sequences. *Bioinformatics* 8, 275–282. doi:10.1093/bioinformatics/8.3.275

Katoh, K., and Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* 30, 772–780. doi:10.1093/molbev/mst010

Knaus, T., Böhmer, W., and Mutti, F. G. (2017). Amine Dehydrogenases: Efficient Biocatalysts for the Reductive Amination of Carbonyl Compounds. *Green. Chem.* 19, 453–463. doi:10.1039/c6gc01987k

Koch, C., Neumann, P., Valerius, O., Feussner, I., and Ficner, R. (2016). Crystal Structure of Alcohol Oxidase from Pichia pastoris. *PLoS.One.* 11–e0149846. doi:10.1371/journal.pone.0149846

Kohls, H., Steffen-Munsberg, F., and Höhne, M. (2014). Recent Achievements in Developing the Biocatalytic Toolbox for Chiral Amine Synthesis. *Curr. Opin. Chem. Biol.* 19, 180–192. doi:10.1016/j.cbpa.2014.02.021

Konagurthu, A. S., Whisstock, J. C., Stuckey, P. J., and Lesk, A. M. (2006). MUSTANG: a Multiple Structural Alignment Algorithm. *Proteins* 64, 559–574. doi:10.1002/prot.20921

Krieger, E., and Vriend, G. (2014). YASARA View-Molecular Graphics for All Devices-From Smartphones to Workstations. *Bioinformatics* 30, 2981–2982. doi:10.1093/bioinformatics/btu426

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol.Biol.Evol.* 35, 1547–1549. doi:10.1093/molbev/msy096

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

Ma, B.-G., Goncearenco, A., and Berezovsky, I. N. (2010). Thermophilic Adaptation of Protein Complexes Inferred from Proteomic Homology Modeling. *Structure* 18, 819–828. doi:10.1016/j.str.2010.04.004

Man, H., Loderer, C., Ansorge-Schumacher, M. B., and Grogan, G. (2014). Structure of NADH-dependent Carbonyl Reductase (CPCR2) from *Candida Parapsilosis* Provides Insight into Mutations that Improve Catalytic Properties. *ChemCatChem* 6, 1103–1111. doi:10.1002/cctc.201300788

Persson, B., Hedlund, J., and Jörnvall, H. (2008). Medium- and Short-Chain Dehydrogenase/reductase Gene and Protein Families. *Cell. Mol. Life Sci.* 65, 3879–3894. doi:10.1007/s00018-008-8587-z

Petruszko, R. (1979). "Nonethanol Substrates of Alcohol Dehydrogenase,". *Biochemistry and Pharmacology of Ethanol*. Editors E. Majchrowicz and E. P. Noble (New York: Plenum Press), Vol. 1, 87–106.

Pirovano, W., Feenstra, K. A., and Heringa, J. (2006). Sequence Comparison by Sequence harmony Identifies Subtype-specific Functional Sites. *Nucleic Acids Res.* 34, 6540–6548. doi:10.1093/nar/gkl901

Riveros-Rosas, H., Julian-Sanchez, A., and Pinã, E. (1997). Enzymology of Ethanol and Acetaldehyde Metabolism in Mammals. *Arch. Med. Res.* 28, 453–471.

Riveros-Rosas, H., Julian-Sanchez, A., Villalobos-Molina, R., Pardo, J. P., and Pina, E. (2003). Diversity, Taxonomy and Evolution of Medium-Chain Dehydrogenase/reductase Superfamily. *Eur. J. Biochem.* 270, 3309–3334. doi:10.1046/j.1432-1033.2003.03704.x

Saitou, N., and Nei, M. (1987). The Neighbor-Joining Method: a New Method for Reconstructing Phylogenetic Trees. *Mol. Biol. Evol.* 4, 406–425. doi:10.1093/oxfordjournals.molbev.a040454

Tan, T. C., Spadiut, O., Wongnate, T., Sucharitakul, J., Krondorfer, I., Sygmund, C., et al. (2013). The 1.6 A crystal Structure of Pyranose Dehydrogenase from Agaricus meleagris Rationalizes Substrate Specificity and Reveals a Flavin Intermediate. *PLoS.One.* 8, e53567. doi:10.1371/journal.pone.0053567

Thungon, P. D., Kakoti, A., Ngashangva, L., and Goswami, P. (2017). Advances in Developing Rapid, Reliable and Portable Detection Systems for Alcohol. *Biosens. Bioelectron.* 97, 83–99. doi:10.1016/j.bios.2017.05.041

Tian, K., and Li, Z. (2020). A Simple Biosystem for the High-Yielding Cascade Conversion of Racemic Alcohols to Enantiopure Amines. *Angew. Chem. Int. Ed.* 59, 21745–21751. doi:10.1002/anie.202009733

Tseliou, V., Knaus, T., Masman, M. F., Corrado, M. L., and Mutti, F. G. (2019a). Generation of Amine Dehydrogenases with Increased Catalytic Performance and Substrate Scope from Epsilon-Deaminating L-Lysine Dehydrogenase. *Nat.Commun.* 10, 3717. doi:10.1038/s41467-019-11509-x

Tseliou, V., Masman, M. F., Böhmer, W., Knaus, T., and Mutti, F. G. (2019b). Mechanistic Insight into the Catalytic Promiscuity of Amine Dehydrogenases: Asymmetric Synthesis of Secondary and Primary Amines. *Chembiochem* 20, 800–812. doi:10.1002/cbic.201800626

UniProt Consortium (2019). UniProt: a Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.* 47, D506–D515. doi:10.1093/nar/gky1049

Vacic, V., Iakoucheva, L. M., and Radivojac, P. (2006). Two Sample Logo: a Graphical Representation of the Differences between Two Sets of Sequence Alignments. *Bioinformatics* 22, 1536–1537. doi:10.1093/bioinformatics/btl151

Vanhanen, S., West, M., Kroon, J. T. M., Lindner, N., Casey, J., Cheng, Q., et al. (2000). A Consensus Sequence for Long-Chain Fatty-Acid Alcohol Oxidases from Candida Identifies a Family of Genes Involved in Lipid ω-Oxidation in Yeast with Homologues in Plants and Bacteria. *J. Biol. Chem.* 275, 4445–4452. doi:10.1074/jbc.275.6.4445

Vanhooke, J. L., Thoden, J. B., Brunhuber, N. M. W., Blanchard, J. S., and Holden, H. M. (1999). Phenylalanine Dehydrogenase fromRhodococcus sp.M4: High-Resolution X-ray Analyses of Inhibitory Ternary Complexes Reveal Key Features in the Oxidative Deamination Mechanism†,‡. *Biochemistry* 38, 2326–2339. doi:10.1021/bi982244q

Wang, S., Nie, Y., Xu, Y., Zhang, R., Ko, T.-P., Huang, C.-H., et al. (2014). Unconserved Substrate-Binding Sites Direct the Stereoselectivity of Medium-Chain Alcohol Dehydrogenase. *Chem. Commun.* 50, 7770–7772. doi:10.1039/c4cc01752h

Wang, X. F., Feng, Z., and Yi, F. L. (2018). X-ray crystal Structure of 2R,3R-Butanediol Dehydrogenase from Bacillus Subtilis. [On-line]. Available: https://www.rcsb.org/structure/6IE0.

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009). Jalview Version 2--a Multiple Sequence Alignment Editor and Analysis Workbench. *Bioinformatics* 25, 1189–1191. doi:10.1093/bioinformatics/btp033

Wu, S., Snajdrova, R., Moore, J. C., Baldenius, K., and Bornscheuer, U. T. (2021). Biocatalysis: Enzymatic Synthesis for Industrial Applications. *Angew. Chem. Int. Ed.* 60, 88–119. doi:10.1002/anie.202006648

Yamamoto, H., Kawada, N., Matsuyama, A., and Kobayashi, Y. (1999). Cloning and Expression inEscherichia Coliof a Gene Coding for a Secondary Alcohol Dehydrogenase fromCandida Parapsilosis. *Biosci. Biotechnol. Biochem.* 63, 1051–1055. doi:10.1271/bbb.63.1051

Yamamoto, H., Matsuyama, A., Kobayashi, Y., and Kawada, N. (1995). Purification and Characterization of (S)-1,3-Butanediol Dehydrogenase fromCandida Parapsilosis. *Biosci. Biotechnol. Biochem.* 59, 1769–1770. doi:10.1271/bbb.59.1769

Yamamoto, H., Matsuyama, A., and Kobayashi, Y. (2002). Synthesis of Ethyl (R)-4-Chloro-3-hydroxybutanoate with RecombinantEscherichia coliCells Expressing (S)-Specific Secondary Alcohol Dehydrogenase. *Biosci. Biotechnol. Biochem.* 66, 481–483. doi:10.1271/bbb.66.481

Yang, Y., Liu, J., and Li, Z. (2014). Engineering of P450pyr Hydroxylase for the Highly Regio- and Enantioselective Subterminal Hydroxylation of Alkanes. *Angew. Chem. Int. Ed.* 53, 3120–3124. doi:10.1002/anie.201311091

Ye, L. J., Toh, H. H., Yang, Y., Adams, J. P., Snajdrova, R., and Li, Z. (2015). Engineering of Amine Dehydrogenase for Asymmetric Reductive Amination of Ketone by Evolving *Rhodococcus* Phenylalanine Dehydrogenase. *ACS Catal.* 5, 1119–1122. doi:10.1021/cs501906r

Zamyatnin, A. A. (1984). Amino Acid, Peptide, and Protein Volume in Solution. *Annu. Rev. Biophys. Bioeng.* 13, 145–165. doi:10.1146/annurev.bb.13.060184.001045

Zamyatnin, A. A. (1972). Protein Volume in Solution. *Prog. Biophys. Mol. Biol.* 24, 107–123. doi:10.1016/0079-6107(72)90005-3

Zeldovich, K. B., Berezovsky, I. N., and Shakhnovich, E. I. (2007). Protein and DNA Sequence Determinants of Thermophilic Adaptation. *Plos Comput. Biol.* 3, e5. doi:10.1371/journal.pcbi.0030005

Zimmermann, L., Stephens, A., Nam, S.-Z., Rau, D., Kübler, J., Lozajic, M., et al. (2018). A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J. Mol. Biol.* 430, 2237–2243. doi:10.1016/j.jmb.2017.12.007

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.