Title

Elzbieta Rembeza[1], Martin KM Engqvist[1*]

[1] Department of Biology and Biological Engineering, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden

* Corresponding author
E-mail: martin.engqvist@chalmers.se

Abstract

250 words

Significance statement

120 words

Introduction

With the steady increase of genetic information deposited to public databases, the proportion of experimentally characterised sequences continues to decline. At the time of writing the UniProt/TrEMBL protein database contains nearly 185 million entries, with only 0.3% of them having been manually annotated and reviewed in the Swiss-Prot database (UniProt Consortium 2019). Furthermore, the experimentally characterized sequence diversity is limited, representing proteins mainly from eukaryotes and model organisms (ref). As current experimental methods for determining protein function cannot keep up with the increase in genomic data, homology-based automated methods form the basis for functional assignment of new proteins (Furnham et al. 2009). Although the automatic annotation pipelines enable processing of vast amounts of newly sequenced data, they cannot predict novel functions and may result in erroneous functional assignments, which later propagate throughout databases (Danchin et al. 2018; Schnoes et al. 2009; Gilks et al. 2005).

Several global initiatives were formed in the past decade to help combat the problem of annotation errors and to develop enabling computational tools that facilitate discovery of novel enzyme functions and improve automated functional predictions. Notable among these initiatives are the Enzyme Function Initiative (Zallot, Oberg, and Gerlt 2019), COMBREX (Roberts et al. 2011), and CAFA (Radivojac et al. 2013).

More recently, several research groups addressed the lack of experimental data using high-throughput methods enabling protein family-wide substrate profiling for hundreds of enzymes. Data generated in such approaches are important for understanding sequence-function relationships in the tested protein families; they have led to the discovery of novel enzymatic activities as well as identified enzymes with diverse physicochemical properties (Bastard et al. 2014; Helbert et al. 2019; Huang et al. 2015; Vanacek et al. 2018). These methods should be further developed and applied to individual enzyme classes in public databases for validating existing functional annotations, as well as for uncovering and correcting annotation errors.

S-2-hydroxyacid oxidases (EC 1.1.3.15) are enzymes which oxidize the hydroxyl group of S-2-hydroxyacids like glycolate or lactate to 2-oxoacids, using oxygen as an electron acceptor (Supplementary figure 1). All characterised enzymes of this class belong to a family of FMN-dependent α-hydroxy acid oxidases/dehydrogenases. Members of this protein family share high structural and functional similarities but differ in the ultimate electron acceptor: oxygen (S-2-hydroxyacid oxidase, EC 1.1.3.15; lactate monooxygenase, EC 1.13.12.4), cytochrome c (flavocytochrome b2, EC 1.1.2.3) or quinone (S-mandelate dehydrogenase, EC 1.1.99.31) (Sukumar et al. 2018; Kean and Karplus 2019; Xia and Mathews 1990). A characteristic feature for S-2-hydroxyacid oxidases is their broad substrate scope *in vitro*, although the physiological substrate for plant and mammalian homologues is mainly glycolate or long chain hydroxyacids (J. M. Jones, Morrell, and Gould 2000; Esser et al. 2014; Dellero et al. 2015), while lactate is the main physiological substrate of bacterial homologues (Umena et al. 2006; Hackenberg et al. 2011).

In this study we provide an overview of the diverse proteins annotated to EC 1.1.3.15 in the BRENDA database (Jeske et al. 2019). We select and experimentally test a diverse subset of these, comprising 122 proteins, for their predicted function. A majority of sequences contain non-canonical protein domains, do not catalyze the canonical reaction, and are therefore wrongly annotated to the enzyme class. Among the misannotated sequences we confirm four alternative enzymatic activities. Finally, a computational analysis of all EC classes in BRENDA reveals that EC 1.1.3.15 is not an outlier in terms of misannotation as large proportion of sequences are annotated to enzyme classes with no similarity to characterised enzymes, and thus are unlikely to perform the predicted function.

Results

Exploration of EC 1.1.3.15 sequence space

EC 1.1.3.15 was chosen for this proof-of-concept study, as it is a medium size, easy to assay class, whose members are of medical and industrial interest, being used for biosensor development (Rassaei et al. 2014; Tsiafoulis, Prodromidis, and Karayannis 2002). To obtain an overview of sequence diversity in EC 1.1.3.15, we downloaded all sequences annotated to this EC in BRENDA 2017.1 and obtained 1058 unique sequences after filtering out partial genes. UniRep embeddings (Alley et al. 2019) were computed for each of these sequences, allowing for alignment-free comparisons, and sequence interrelatedness was visualized in an MDS plot, where a smaller distance indicates higher relatedness (Figure1, Supplementary figure 2). 17 of these sequences are characterised enzymes: either listed in BRENDA (Jeske et al. 2019) as experimentally tested or in SwissProt (UniProt Consortium 2019) as having experimental evidence at protein level. Over 90% of the enzymes annotated to this enzyme class are of bacterial origin, nearly 6% of eukaryotic and 2.6% of archaeal (Figure 1a). Strikingly, 14 out of 17 characterised enzymes are of eukaryotic origin, showing a clear over-representation. The characterized sequences also cluster close together, indicating that the experimentally tested sequences diversity in EC 1.1.3.15 is limited.

We next determined similarity of each sequence in EC 1.1.3.15 to the closest characterised S-2-hydroxyacid oxidase in terms of sequence identity and domain architecture. Most sequences have little similarity with the characterized ones; 79% of sequences annotated as 1.1.3.15 share less than 25% sequence identity with the closest biochemically characterised sequence, which are mainly eukaryotic (Figure 1b, Supplementary figure 3). Furthermore, only 22.5% of the 1058 sequences are predicted to contain the FMN-dependent dehydrogenase domain (FMN_dh, PF01070) which is canonical for known 2-hydroxy acid oxidases (Figure 1c). The majority of sequences were predicted to contain non-canonical domains, such as FAD binding domains characteristic for FAD-dependant oxidoreductases (PF01266, PF01565, PF02913), as well as a cysteine rich domain (PF02754) and 2Fe-2S binding domain (PF04324). [glcE] Many of the sequences with non-canonical domains form distinct clusters (Figure 1c). An analysis of similarity between these domain clusters showed that the average sequence identity to the canonical FMN-dependent dehydrogenase domain cluster is below 16% for all clusters. An all versus all comparison revealed that no two clusters share more than 21% average sequence identity, while the identity of sequences within clusters ranges between 33% and 55% (Figure 1d).

This analysis clearly shows that the enzyme class EC 1.1.3.15 contains a set of very diverse protein sequences, the majority of which have low identity to the previously characterised enzymes, and also lack protein domains characteristic of S-2-hydroxy acid oxidases.

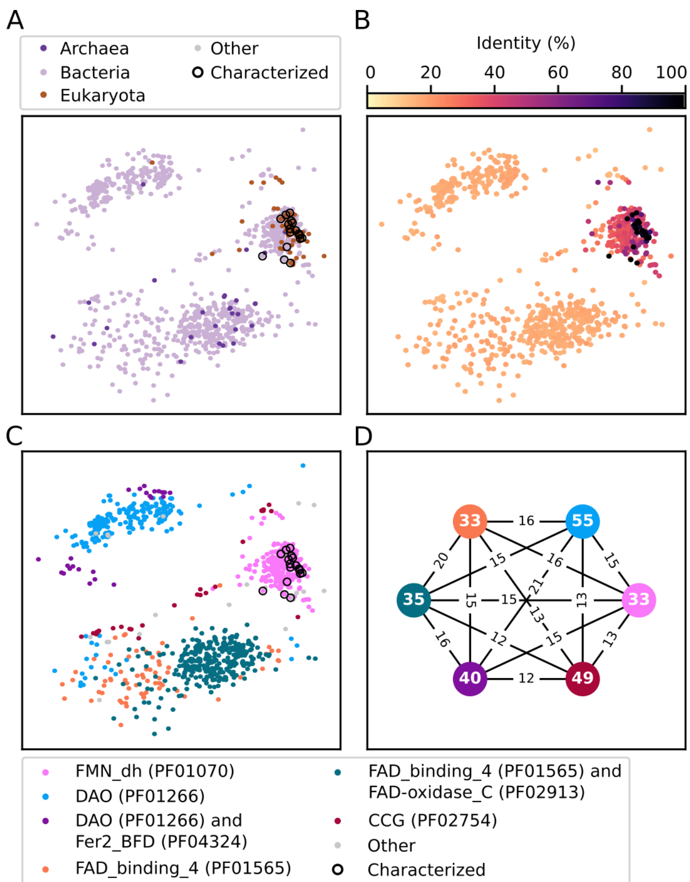

Figure 1 | Sequence space of sequences annotated to EC 1.1.3.15. Sequences listed in BRENDA and SwissProt as experimentally tested are encircled **(A)** Taxonomic origin of sequences. **(B)** Percentage of sequence identity to the closest experimentally tested S-2-hydroxyacid oxidase. **(C)** Predicted Pfam domain architecture. **(D)** The mean sequence identity between and within domain clusters. Pfam protein domains: FMN_dh (PF01070) - FMN-dependent dehydrogenase, DAO (PF01266) - FAD dependent oxidoreductase, Fer2_BFD (PF04324) - BFD-like [2Fe-2S] binding domain, FAD_binding_4 (PF01565) - FAD binding domain, FAD-oxidase_C (PF02913) - FAD linked oxidases C-terminal domain, CCG (PF02754) - cysteine-rich domain.

Experimental characterisation of EC 1.1.3.15

Due to the large diversity of sequences annotated to EC 1.1.3.15 we wished to experimentally validate their predicted activity. A total of 122 genes throughout the sequence space of the enzyme class were selected (Supplementary figure 4a), synthesized, cloned and recombinantly expressed in *Escherichia coli* in a high throughput set up. Out of the 122 proteins, 65 were in soluble state (53%), with archaeal and eukaryotic proteins being proportionally less soluble than bacterial proteins (Supplementary figure 4b). Despite representing only half of the sequences chosen for experimental characterisation, the soluble proteins were still distributed throughout the sequence space of EC 1.1.3.15 (Supplementary figure 4a). The 65 soluble proteins were tested for S-2-hydroxy acid oxidase activity in an Amplex Red peroxide detection assay with a set of six 2-hydroxy acids: glycolate, lactate, 2-hydroxyoctanoate, 2-hydroxydecanoate, mandelate, 2-hydroxyglutarate (Supplementary figure 5).

Characterisation of proteins carrying the canonical FMN-dh domain

We first investigated 24 proteins representing a cluster of 230 sequences containing the FMN_dh domain; these have the highest sequence identity to previously characterised 2-hydroxy acid oxidases (Figure 1c, Figure 2a). Among them 14 proteins were active with a broad substrate range, as is characteristic for enzymes in EC 1.1.3.15, while 10 proteins were inactive. Bacterial sequences in the cluster were predominantly active with lactate, medium chain and aromatic 2-hydroxy acids, whereas the two active eukaryotic enzymes showed the highest activity with glycolate and lactate.

We next analyzed whether the 24 investigated proteins contain the seven conserved amino acid residues involved in catalysis and substrate binding (Dellero et al. 2015), both using a multiple sequence alignment and protein structure analysis (Figure 2a and b). In 12 of the 14 active proteins all seven residues are conserved (Figure 2a), whereas 8 of the 10 inactive proteins lack at least one of the conserved residues. Presence of the seven conserved amino acids is thus a strong - but not absolute - indication of S-2-hydroxyacid oxidase activity.

The seven active site residues are, however, conserved not only in S-2-hydroxyacid oxidases, but also among all the members of FMN-dependant S-2-hydroxyacid oxidase/dehydrogenase family (Kean and Karplus 2019). We therefore looked for sequence motifs indicating the presence of other family members in our selection (Figure 2c). Two of the proteins (B8MKR3 and B8MMC0 from *Talaromyces stipitatus*) in the cluster contain a heme binding domain (PF00173) characteristic for flavocytochrome b2 L-lactate dehydrogenase (EC 1.1.2.3) (Xia and Mathews 1990) (Figure 2a, Supplementary figure 6). The two proteins were tested *in vitro* for their ability to reduce cytochrome c, a physiological electron acceptor of flavocytochrome b2 L-lactate dehydrogenase. Indeed, the B8MKR3 protein displayed the cytochrome b2 L-lactate dehydrogenase activity (Supplementary figure 7). Additionally, three other proteins (E6SCX5 from *Intrasporangium calvum*, C9Y9E7 from a *Curvibacter* species and W6W585 from *Rhizobium sp. CF080*) contain a longer stretch in loop 4 characteristic for S-mandelate dehydrogenase (EC 1.1.99.31) and L-lactate 2-monooxygenase (EC 1.13.12.4)

(Kean and Karplus 2019; Sukumar et al. 2018) (Figure 2a, Supplementary figure 6). As seen in our Amplex Red assay the three proteins display a higher activity with mandelate than other proteins in the cluster, suggesting their native function may be as S-mandelate dehydrogenases, but further experiments are needed to determine this.

Out of the 230 members of the FMN_dh cluster - with high sequence identity to previously characterized EC 1.1.3.15 enzymes - a total of 6 proteins (2.6%) are predicted to contain a heme binding domain and 50 (22%) contain a longer stretch in loop4, indicating that those sequences might be misannotated and would be better placed in other EC classes. However, a thorough biochemical and genetic characterisation of such enzymes is needed to test this hypothesis.

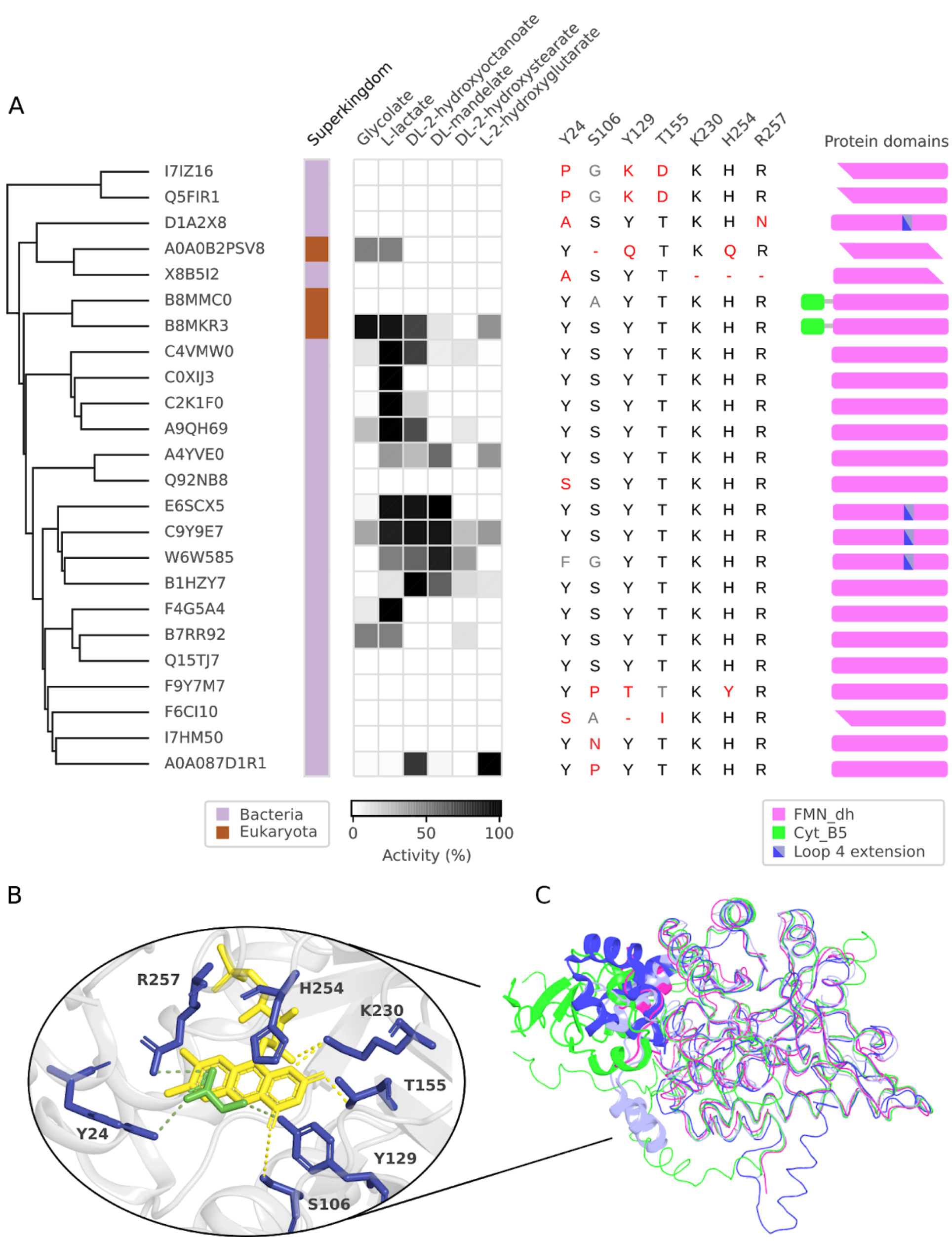

Figure 2 | Characterisation of protein cluster with high sequence identity to previously characterised S-2-hydroxyacid oxidases. (A) Activity screen and protein characteristics. [what the dendrograms are]. Superkingdoms: light purple - Bacteria, brown – Eukaryotes. Recorded activities are marked with squares, for proteins active with more than one substrate, the substrate preference is shaded with the highest activity for each enzyme scaled to 100%. Listed amino acids correspond to conserved residues in a glycolate oxidase from *S. oleracea.* The cartoons represent predicted domain and motif composition of the sequences, based on Pfam search. Domains lacking full Pfam alignment are represented with a sharp edge. FMN-binding domain (FMN_dh, PF01070) is marked in magenta, cytochrome b5-like heme binding domain (Cyt_B5, PF00173) is marked in green, and a prolonged stretch in loop4 is marked in blue. **(B)** Conserved amino acids of the active site of S-2-hydroxyacid oxidase mapped on a structure of glycolate oxidase from *S. oleracea* (PDB: 1GOX). Conserved residues are marked in blue, the FMN cofactor is marked in yellow, and the glycolate substrate in green. **(C)** Superimposed structures of the representatives of FMN-dependant 2-hydroxyacid oxidase/dehydrogenase family with their distinct motifs represented in a cartoon form: glycolate oxidase (magenta, PDB 1GOX), flavocytochrome b2 (green, PDB 1FCB), mandelate dehydrogenase (light blue, PDB 6BFG), lactate 2-monooxygenase (dark blue, PDB 6DVH).

Characterisation of proteins carrying non-canonical domains

Next, we investigated the activity of 41 proteins not containing the canonical FMN-dh domain (Figure 1c), yet representing a full 78% of all sequences annotated to EC 1.1.3.15 in BRENDA. These proteins have only low sequence identity with previously characterised S-2-hydroxyacid oxidases (Figure 1b and d).

Out of the 41 proteins twelve come from the cluster predicted to contain a single FAD dependent oxidoreductase domain (DAO, PF01266). Six of the twelve solely oxidized the substrate L-2-hydroxyglutarate in the *in vitro* assay (Figure 3a). This narrow substrate scope is atypical for the previously known broad substrate-range EC 1.1.3.15 enzymes, which indicates an alternative native function of these proteins. Our findings are supported by those of a recent publication where activity of an *E. coli* homologue of the 6 DAO-containing proteins was described as L-2-hydroxyglutarate dehydrogenase (EC 1.1.99.2) (Knorr et al. 2018). As the Amplex Red activity assay used in our activity screen is designed to capture oxidase activity via hydrogen peroxide detection, we may have detected a low level of non-physiological oxidase activity of the 6 L-2-hydroxyglutarate dehydrogenases (see further discussion on the AR assay a few paragraphs below).

Remaining 29 sequences of the “non-canonical” clusters - containing either a BFD-like [2Fe-2S] binding domain (Figure 3a), or a FAD linked oxidases C-terminal domain, either alone or combined with a cysteine-rich domain (Figure 3b) - were either inactive or did not display consistent substrate preferences (Figure 3a and b). We hypothesized that due to the non-canonical domain architecture and low sequence identity to characterized enzymes, these proteins may catalyze reactions different from the ones initially tested. By searching database information regarding the predicted Pfam (El-Gebali et al. 2019) domains and combining this information with orthology-based annotations and literature search, we found that some of these sequences are similar to dehydrogenases operating on four distinct substrates: glycerol-3-phosphate, glycolate, D-lactate and D-2-hydroxyglutarate dehydrogenase.

In order to test whether the proteins in our selection catalyze these alternate reactions, we successfully purified 22 out of 29 proteins and screened them for the expected dehydrogenase activities with a set of common electron acceptors: nicotinamide adenine dinucleotide (NAD), nicotinamide adenine dinucleotide phosphate (NADP), the redox dye 2,6-Dichlorophenolindophenol (DCPIP), as well as the hydrogen peroxide probe Amplex Red (AR), and in selected cases cytochrome c (Supplementary figure 8). When screened with DCPIP and AR, one protein was found to be active with glycerol-3-phosphate as a substrate (A0A0R3K2G2 from *Caloramator mitchellensis*), one with D-lactate (D4MUV9 from *Anaerostipes hadrus*) and one with D-2-hydroxyglutarate (A0A077SBA9 from *Xanthomonas campestris*). Additionally, three proteins (A0A0U5JSS4 from a *Clostridium* species, D4XIR1 from *Achromobacter piechaudii*, Q5WIP4 from *Bacillus clausii*) were active with each of the three substrates only in the AR screen (Supplementary figure 8). None of the proteins were active with the electron acceptors NAD, NADP, or cytochrome c.

The fact that some of the tested enzymes show activity with both AR and DCPIP is counter-intuitive as AR is a H_2O_2-dependent reporter, indicating that molecular oxygen is the electron acceptor, whereas DCPIP accepts electrons directly. Comparing standard curves of the two reporter molecules DCPIP and resorufin (the AR reaction product) revealed that the AR assay is several orders of magnitude more sensitive than DCPIP, on a molar basis (Supplementary figure 9a). We then carried out a direct comparison of enzyme activity in four purified enzymes using the DCPIP and AR assays. While the AR-dependent assay clearly gave the strongest signal, the enzymes displayed fifty to one hundred times higher catalytic rates in the DCPIP-based one (Supplementary figure 9b). Dehydrogenase activity is thus the prevalent one for the tested enzymes, although we were able to capture their trace oxidase activity.

Overall, our screen of the non-canonical clusters revealed their erroneous annotation as EC 1.1.3.15, and we found four alternative activities among those sequences: L-2-hydroxyglutarate dehydrogenase, D-2-hydroxyglutarate dehydrogenase, D-lactate dehydrogenase, and glycerol-3-phosphate dehydrogenase. Four representatives with the alternative activities were chosen for further characterization (Figure 3a and b, in bold); they were expressed, purified (Supplementary figure 10A), assayed at 25 °C and their kinetic parameters calculated (Table 1, Supplementary figure 10B). Three of the four enzymes (D4MUV9, A0A077SBA9, S2DJ52) had substrate affinities in the micromolar range and high catalytic rates, strengthening the possibility that these may be the natural substrates. Additionally, based on reports of a homologous protein (Guo et al. 2018), the protein A0A077SBA9 was screened and proved to show modest side activity with D-malate. The fourth enzyme, A0A0R3K2G2, showed affinity for glycerol-3-phosphate in the low millimolar range, but with catalytic rates approximately 20-fold lower than the other enzymes. Since this protein comes from a thermophilic bacterium *Caloramator mitchellensis* whose optimal growth temperature is 55 °C, we speculate that the catalytic rate would be higher at higher experimental temperatures. Taken together, our results indicate that proteins which do not contain the canonical FMN-dh domain, which represent 78% of all proteins annotated to EC 1.1.3.15 in BRENDA, likely have *in vitro* catalytic activities that do not match their current EC classification.

Table 1. Kinetic parameters of selected proteins with functions distinct from L-2-hydroxyglutarate oxidase. Values represent mean averages (+/- standard error of mean; n = 3).

Enzyme	Substrate	K_M [mM]	V_{max} [U/mg]	k_{cat} [s^{-1}]
D4MUV9	D-lactate	0.396 +/- 0.042	6.016 +/- 0.168	5.180
A0A077SBA9	D-2-hydroxyglutarate	0.082 +/- 0.009	7.008 +/- 0.197	5.957
	D-malate	5.031 +/- 1.381	0.046 +/- 0.005	0.039
S2DJ52	L-2-hydroxyglutarate	0.221 +/- 0.015	5.072 +/- 0.091	3.719
A0A0R3K2G2	glycerol-3-phosphate	1.972 +/- 0.233	0.273 +/- 0.009	0.242

Analysing annotation error in the BRENDA database

Biological databases are dynamic by nature and receive regular updates with new experimental information as well as additional proteins from sequenced genomes genomes. We therefore wanted to investigate how the annotations to EC 1.1.3.15 changed over time.

In our analysis we compared predicted Pfam domains of sequences annotated to the class in BRENDA 2017.1 and BRENDA 2019.2 (Figure 3c). Over the course of 2.5 years, representing five database versions, the enzyme class grew markedly, by 601 sequences to 1659 (excluding redundant and partial sequences). However, the number of sequences containing the canonical FMN-dh domain actually decreased by 11, whereas the newly added sequences are part of clusters containing "non-canonical" protein domains. The most striking rise in sequences in this time period, from 24 to 220 sequences, appeared in the cluster shown by us to contain proteins displaying glycerol-3-phosphate dehydrogenase activity (Pfam domains DAO and Fer2_BFD) *in vitro* as well as that containing the L-2-hydroxyglutarate dehydrogenases (Pfam domain DAO), which rose from 379 to 650 sequences.

This comparison clearly shows that, in the EC 1.1.3.15 enzyme class, the misannotations from old database versions were perpetuated to newly added homologous sequences. Based on the number of sequences lacking the canonical domain architecture alone (absence of the canonical FMN dehydrogenase domain) we estimate that in 2017 at least 78% of sequences in EC 1.1.3.15 are unlikely to catalyze the predicted reaction, while in 2019 this number grew to 87%.

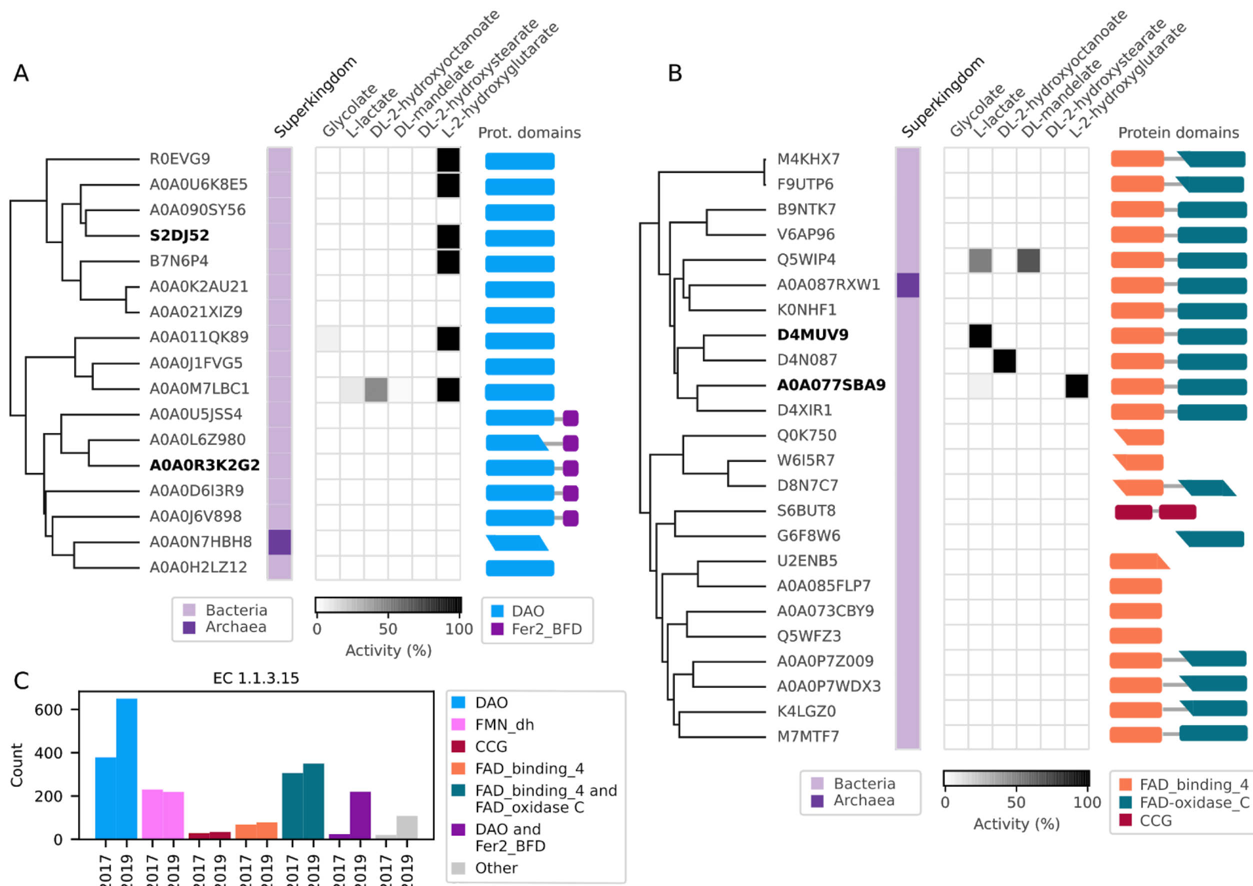

Figure 3 | **Characterisation of protein clusters with low sequence identity to previously characterised S-2-hydroxyacid oxidases.** [what the dendrograms are]. Superkingdoms: light purple - Bacteria, dark purple - Archaea. Activities are marked with squares, for proteins active with more than one substrate, the substrate preference is shaded. The cartoons represent predicted domain and motif composition of the sequences, based on Pfam search. Domains lacking full Pfam alignment are represented with a sharp edge. Proteins with alternative activities chosen for kinetic characterisation are marked in bold. **(A)** Characterisation of protein clusters containing DAO domain. FAD dependent oxidoreductase domain (DAO, PF01266) is marked in blue, BFD-like [2Fe-2S] binding domain (Fer2_BFD, PF04324) is marked in purple. **(B)** Characterisation of remaining protein clusters. FAD binding domain (FAD_binding_4, PF01565) is marked in orange, FAD linked oxidases C-terminal domain (FAD-oxidase_C, PF02913) is marked in green, cysteine rich domain (CCG, PF02754) is marked in red. **(C)** Comparison of predicted Pfam domains of sequences annotated to EC 1.1.3.15 in BRENDA version 2017.1 and 2019.2

Exploration of functional annotations in other enzyme classes

In our initial analysis of EC 1.1.3.15 we observed that enzymes from eukaryotes had been disproportionately studied and that a large proportion of sequences annotated to the class shared little similarity with them (Figure 1). We next asked whether EC 1.1.3.15 is a special case, or whether these observations constitute a trend across all of BRENDA. To answer this question we first downloaded all protein sequences from BRENDA 2019.2 and determined which of these have experimental evidence in BRENDA or SwissProt. We found 30 574 unique identifiers with experimental evidence in SwissProt and 31 287 in BRENDA, only 11 498 of which were overlapping between the two sources. Next, we determined, for each EC class in BRENDA, the degree of identity between each uncharacterized sequence with the most similar characterized one. To decrease the effect of a large number of similar sequences from repeated sequencing of model organisms (ref?) we clustered the sequences at 90% using CD-HIT (Li and Godzik 2006) and carried out the subsequent analysis using cluster representatives only. As in EC 1.1.3.15 (Figure 1), this global analysis shows that the overwhelming majority of sequences in BRENDA are bacterial (Figure 4a), whereas the majority of experimentally characterized enzymes are eukaryotic (Figure 4b). Furthermore, most enzyme classes have only a small number of characterized enzymes (Figure 4c), indicating that the sequence diversity explored within each EC class is limited.

To analyze the similarity of uncharacterised sequences to characterised ones we computed, for each EC class, the sequence identity of each cluster representative to the closest characterised enzyme. This analysis is analogous to the one carried out for EC 1.1.3.15 (Figure 1b). The results for all EC classes were aggregated and are presented in Figure 3. In all three superkingdoms the identities roughly follow a normal distribution with a mean below 50% identity (Figure 3d, e and f). Peaks at 0% represent enzymes for which no characterized homolog is known, and peaks at 100% represent enzymes that have themselves been characterized. We also note peaks around 18% identity, these represent the average pairwise identity of two randomly selected sequences within an EC class (Supplementary figure 12). Strikingly, in each of the superkingdoms more almost one fifth of sequences share less than 25% pairwise sequence identity with the closest characterized enzyme - within their own EC class. Such sequences are likely to be incorrectly annotated to a given EC, considering that this is well below the level where function can be confidently transferred between homologous proteins (twilight ref). Many such low-homology sequences are annotated even to ostensibly well-characterised enzyme classes with industrially relevant activities (Table 2).

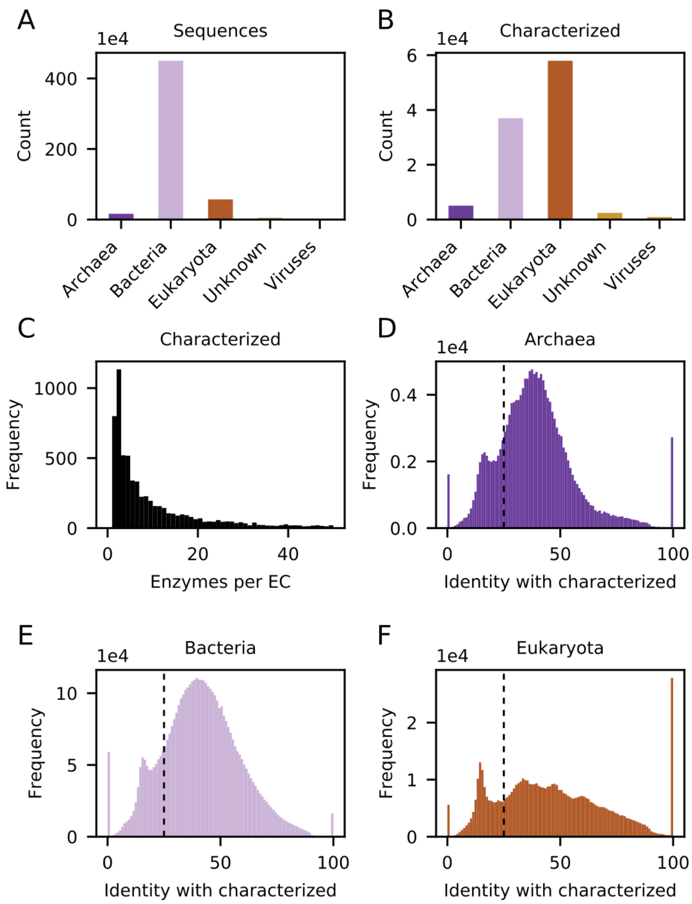

Figure 4 | Exploration of functional annotation throughout all BRENDA enzyme classes. (A) The total number of protein sequences (after clustering at 90% identity) annotated to EC classes in BRENDA. **(B)** The total number of experimentally characterized enzymes. **(C)** Histogram showing the number of characterized enzymes per EC class (bin size of nnn). Histograms showing the distribution of sequence identities between uncharacterized proteins and their closest characterised enzyme for Archaea **(D)**, Bacteria **(E)**, and Eukaryota **(F)** (with a bin size of nnn).

Table 2 | Overview of annotation to enzyme classes of industrial interest.

EC	**Name**	**%id < 25%***	**Number of characterised proteins****	**Applications** (Singh et al. 2016)
3.1.1.3	lipase	**54.7**	141	detergent, leather processing, pharmaceutical synthesis, degradation of crude oils and plastics
3.1.1.1	carboxylesterase	**47.6**	106	degradation of plastics
3.2.1.4	cellulase	**30.6**	191	pulp and paper processing, detergent
3.2.1.8	xylanase	**29.9**	210	animal feed processing, pulp and paper processing

3.2.1.1	alpha amylase	**23.9**	87	flour adjustment, detergent, leather processing
3.1.1.74	cutinase	**10.2**	28	detergent, degradation of plastics

* Percentage of sequences in the EC with less than 25% identity to the closest characterised enzyme of the EC

** Proteins listed as characterised in BRENDA DB and/or with “experimental evidence at protein level” label in SwissProt

Discussion

In this study we present the first large-scale experimental investigation of sequence space to explore misannotation in a single enzyme class. In contrast to the previous studies investigating annotation errors (Schnoes et al. 2009; C. E. Jones, Brown, and Baumann 2007), our setup allowed us not only to estimate the error, but also to examine alternative functions of the misannotated sequences. Our experimental approach to the misannotation problem comes with a drawback of limited scope, as we describe in detail only one enzyme class, whereas bioinformatic approaches allow for much broader analysis. However, we argue that our setup is ideal for understudied enzyme classes, and protein families for which experimental evidence is scarce.

The most comprehensive misannotation study so far provided an overview of annotation error in 37 enzyme families, yet all the families were well-studied and no additional experimental evidence was required to conduct it (Schnoes et al. 2009). In the work by Schnoes et al. only 3% of all sequences were considered misannotated due to the lack of similarity to the golden standard of a superfamily. In our study we show that this number is likely much higher now. Although we did not explore all possible causes of misannotation, we show that 17.8% of all sequences annotated in BRENDA share less than 25% sequence identity to the nearest characterised enzyme of the class, and thus are unlikely to perform the predicted function. In the example of EC 1.1.3.15 we show that this number can reach as high as 87%, with the degree of misannotation increasing over time.

Accumulation of annotation errors is a direct effect of how the genomes are annotated (Gilks et al. 2005). New entries to a database are usually annotated based on a “majority rule” - similarity to already existing entries, with little concern as to what the previous annotations were based on. This can result in a sequence being annotated not based on similarity to the closest characterised sequences, but on similarity to the largest number of already annotated sequences – whether the annotations were carried correctly or not (Richardson and Watson 2013; Danchin et al. 2018). In our work we present a tangible consequence of such approach; over the span of 2.5 years and five BRENDA database versions accumulation – rather than correction – of annotation errors to EC 1.1.3.15 occurred. Usually the true causes of erroneous annotations are difficult to track and even more difficult to correct, as the correction of deposited genomes in archival databases can only be performed by original authors (Salzberg 2007). Secondary protein databases, such as UniProt or BRENDA, welcome users’ corrections, however, it is uncertain to what extent those options are actively used by the community and result in correction of annotations. In our work we chose to investigate functional annotations to

BRENDA database (Jeske et al. 2019) as it is the premier database linking protein entries with biochemical data, and due to its status as an ELIXIR core data resource (https://elixir-europe.org/platforms/data/core-data-resources), but we expect similar levels of annotation error in all major databases.

The most reliable gene annotations are the ones based on similarity to already characterised gene products, however, not all biochemically characterised proteins are recorded in protein databases and can be used for annotation transfer. In our study we characterised four proteins annotated to EC 1.1.3.15 with alternative activities, and in all cases after a literature search we found articles describing homologous proteins with the same activities (Knorr et al. 2018; Koga et al. 2019; Weghoff, Bertsch, and Müller 2015; Guo et al. 2018). Only one article postulated for annotation transfer (Knorr et al. 2018) which resulted in a recent re-annotation of the protein in UniProt, whereas the remaining proteins are still not recorded in protein databases as being experimentally tested. Initiatives such as COMBREX DB, a database of experimentally validated gene annotations (Chang et al. 2016), or STRENDA, a guideline of standards for reporting enzymology data (Tipton et al. 2014; Swainston et al. 2018) could help to solve the problem, but only if the whole scientific community adopts these standards. As a response to this issue, the journal *Biochemistry* recently called authors to include accession IDs for all proteins experimentally characterized in their manuscripts (Gerlt 2018), a requirement that should certainly be adopted by other journals. We believe that a structured way of registering proteins characterised in high-throughput experiments should also be developed, as though the depth of protein characterisation in such approaches is limited, they can provide an excellent overview of substrate scope of a large number of proteins.

Incorrect gene annotations that accumulate over time might have serious consequences for all the areas of bioscience basing their investigations on accurate annotations, such as systems biology (Griesemer et al. 2018) or metabolic engineering (Erb 2019). Sequence similarity is by no means a perfect determinant of functional transfer, however, our study shows that a large percentage of enzymes are annotated to an EC with almost no sequence similarity to experimentally characterised proteins. We believe that the identity to the nearest characterised sequence combined with prediction of domain architecture should be a vital checkpoint in functional annotations of proteins. Although this approach might result in less densely annotated genomes, their overall quality will be of much higher standards.

Materials and methods

TODO

Modify code to extract UniRep representations, instead of supplying them (then I can provide all the files in the GitHub repo, except for the very large BRENDA files, which we can host on Zenodo)
Take out our other activity data (non 1.1.3.15) from the experiment files.
Clean up wsvg

Clean up orgtools
Clean up brenparse
Clean up unirep50
Add line to notebooks that install the above libraries
Upload key raw datafiles to Zenodo
Add code that downloads raw datafiles to repo
Final cleanup of the 1.1.3.15 repo + README

EC 1.1.3.15 sequence space analysis

All protein sequences from BRENDA (https://www.brenda-enzymes.org/, version 2017.1) were downloaded and their full UniRep embeddings (Alley et al. 2019), of 5700 values, were computed. Identical sequences were de-duplicated and multidimensional scaling (MDS) was carried out on the remaining representations using the builtin function in Scikit-learn (Pedregosa et al. 2011) to decrease the dimensionality of this representation to two, thus allowing visualization as a scatterplot (Figure 1). Taxonomic information for each sequence was obtained by searching for the source organisms name in NCBI Taxonomy resource (https://www.ncbi.nlm.nih.gov/taxonomy). Sequences considered as “characterized” were obtained from UniProtKB/Swiss-Prot (https://www.uniprot.org/) as well as from BRENDA. Specifically, all protein identifiers from UniProtKB/Swiss-Prot (version 2020_02) annotated as belonging to EC 1.1.3.15 and labelled with “Evidence at protein level” were used, as well as those occurring in the “Organism” table of the EC 1.1.3.15 html page in BRENDA (version 2019.1). Pairwise sequence alignments were carried out, using MUSCLE (Edgar 2004), between all 1.1.3.15 sequences. For each sequence the maximum identity to a characterized one was retained (Figure 1b). Pfam protein domain information for each sequence was obtained from UniProtKB. For the domain architectures specified in Figure 1d the arithmetic mean of all pairwise identities was calculated, within each architecture, as well as between architectures.

Sequence selection for experimental testing

Protein sequences from all EC classes designated as being oxidoreductases acting on hydroxyl groups with oxygen as an electron acceptor (EC 1.1.3.-) were downloaded from BRENDA (version 2017.1) and processed as outlined below, but only sequences from 1.1.3.15 were tested here, the others being reserved for future work. To improve the quality of subsequent alignments, sequences shorter than 200 amino acids (61 total for EC 1.1.3.15) and longer than 580 (31 total for EC 1.1.3.15) were removed, as well as sequences with “X” in them (7 total for EC 1.1.3.15). An all versus all BLAST was carried out using plastp from BLAST+ (Camacho et al. 2009) with standard settings, followed by clustering using the MCL algorithm (Enright, Van Dongen, and Ouzounis 2002) with standard settings, except for the inflation parameter -I, which was set to 1.4. This resulted in 17 clusters. A multiple-sequence alignment was created for each cluster using MUSCLE (Edgar 2004). The Shannon entropy was calculated for each multiple sequence alignment, and for each cluster sequences were iteratively selected so as each newly chosen sequence maximally increases the mutual information explained within each cluster.

This iterative sequence selection was continued until 85% of the information in each cluster had been explained. In the first round a total of nnn EC 1.1.3.15 sequences were selected from the 17 clusters. After initial biochemical assays had been carried out on these sequences, yielding information of which proteins were not active, a second selection of nnn sequences was made, for a total of nnn.

Cloning, expression of sequences and protein purification

Generated sequences were synthesized, cloned into the pET21a vector and sequenced-verified by Twist Bioscience. Between a sequence and vector, a C-terminal linker was added (AAALEHHHH), which in combination with six histidines from an expression vector resulted in a deca-His-tag for improved protein purification. High throughput expression, lysis and, when necessary, purification was carried out according to the published protocol (Repecka et al. 2019). Briefly, expression was carried in *E. coli* BL21(DE3) cells, in 96-well deep well plates, in 1 ml autoinduction TB (Foremedium). After cell lysis, cells were spun down and supernatants analysed by SDS-PAGE followed by Coomassie staining (InstantBlues, Expedeon). Each sequence was expressed three times, a sequence was scored as soluble when the corresponding band was present on a gel in at least two expressions. Soluble fraction of the lysate was used for the screen of S-2-hydroxyacid oxidase activity, whereas affinity-purified proteins were used for the dehydrogenase activity screen and kinetic parameters calculation.

Activity assays

To screen for the S-2-hydroxyacid oxidase activity, lysates of soluble proteins were assayed in the Amplex Red hydrogen peroxide detection assay (Fisher Scientific) with a selection of 2-hydroxyacids: glycolate, L-lactate, DL-2-hydroxyoctanoate, DL-2-hydroxyoctadecanoate, DL-mandelate, L-2-hydroxyglutarate. Each protein was assayed three times and was considered a hit if it was scored as soluble and active at least twice. 1 µl of soluble fraction of the lysate after protein expression was added to a reaction mixture containing 20 mM HEPES pH 7.4, 50 µM Amplex Red (Fisher Scientific), 0.1 U/ml HRP and 1 mM of an appropriate substrate. Final reaction volume was 20 µl, the assay was performed in black 384-well low volume plates (Greiner). After 30 minutes of incubation in the dark, the endpoint measurements were performed with an excitation filter of 544 nm and emission filter of 590 nm in a BMG Labtech FLUOstar Omega microplate reader. Each reaction was performed in triplicates. Values for the unspecific activity of no substrate controls were subtracted from the other values. *E. coli* lysate from cells expressing BSA protein was used as a control to establish a limit of detection of the assay ($mean_{BSA} + 4*SD_{BSA}$).

For the dehydrogenase activity screening and kinetic characterisation, proteins were purified by affinity purification, and assayed with a range of substrates and electron acceptors. 1 µl of purified protein was added to a reaction mixture containing 20 mM HEPES pH 7.4, 2 mM of substrate and electron acceptor. L-lactate (cytochrome) dehydrogenase activity was tested with 0.1 mM cytochrome c as electron acceptor. Glycerol-3-phosphate dehydrogenase activity was tested with following electron acceptors: 0.2 mM DCPIP + 3 mM PMS, 50 µM Amplex Red +

0.1U/ml HRP, 1mM NAD, 1mM NADP. 2-hydroxyacid dehydrogenase activity was tested with all the above electron acceptors, with the addition of 0.15 mM cytochrome c. Activity was measured in triplicates every 30 seconds over 15 minutes at 340 nm in case of NAD and NADP, at 600 nm in case of DCPIP/PMS, at 550 nm in case of cytochrome c, and with excitation/emission filter of 544 nm/590 nm in case of Amplex Red/HRP. Unspecific reduction of electron acceptor was monitored in no substrate controls, and the values obtained were subtracted from the other values.

The kinetic values for four chosen proteins were determined at 25 °C with DCPIP + PMS as electron acceptor and a varied range of substrate concentrations. Protein concentrations used for the assays: 60 nM D4MUV9, 50 nM A0A077SBA9 with D-2-hydroxyglutarate, 1.3 μM A0A077SBA9 with D-malate, 25 nM S2DJ52, 660 nM A0A0R3K2G2. Activities were calculated using the extinction coefficient of DCPIP at 600 nm (20.7 $mM^{-1}cm^{-1}$).

Comparison of DCPIP and AR reaction rates was carried for the four characterised proteins. Reactions rates for a chosen protein and substrate concentrations were performed for both electron acceptors.

EC 1.1.3.15 annotation over time

All EC 1.1.3.15 sequences were downloaded from two BRENDA versions, differing by 2.5 years in their publication (versions 2017.1 and 2019.2). Identical sequences in each database version were de-duplicated, resulting in 1058 sequences from 2017.1 and 1659 sequences from 2019.2. Pfam domains for these sequences were obtained by querying UniProt using the protein identifiers, and mining the resulting page for domain data. The frequency of each domain was subsequently computed.

Exploration of annotation quality throughout enzyme classes

A list of UniProt identifiers for enzymes considered “characterized” was compiled from SwissProt and BRENDA as described in the first Methods section. Protein sequences from all EC classes were downloaded from BRENDA (version 2019.2). Within each EC class sequences were clustered to 90% identity using CD-HIT (Li and Godzik 2006) with standard settings and a word size of 5. Cluster representatives were retained for subsequent analysis. Since the clustering had resulted in some “characterized” sequences to be removed (they were not cluster representatives) these were added back. For every cluster representative within each EC class the sequence identity to the closest characterized sequence (within that class) was computed. First, an alignment-free measure of similarity was obtained using the alfpy package (Zielezinski et al. 2017) by computing count-based k-tuples with word size of 3 and Normalized Google Similarity (Choi and Rashid 2008) as a distance measure. For each uncharacterized-characterized pair with highest k-tuple-based similarity pairwise sequence alignments were created using MUSCLE and the sequence identities calculated. These are the identities reported. The superkingdom of the source organism was obtained for each organism, firstly by matching the organism name with the NCBI-Taxonomy database, and secondly by querying UniProt using the protein identifiers.

Software

Full miniconda environment available together with analysis scripts
Include singularity image?
(https://github.com/EngqvistLab/analyze_1.1.3.15)
Taxonomy scripts (https://github.com/EngqvistLab/orgtools)
Parsing BRENDA (https://github.com/EngqvistLab/brenparse)
My UniRep version (https://github.com/EngqvistLab/UniRep50)
Drawing dendrograms and activity data (https://github.com/mengqvist/wsvg)
Python version 3.7
Biopython (Cock et al. 2009)
MUSCLE
CD-HIT (Li and Godzik 2006)
MCL (version 14-137)
BLAST+
Alfpy (Zielezinski et al. 2017)
UniRep
Scikit-learn
Many more… look through scripts..

Data availability

Zenodo upload

- 1.1.3.15 fasta files for both BRENDA versions used
- 1.1.3.15 clusters?
- Data regarding which sequences were soluble and which not (1.1.3.15_BRENDA_2017_1_sequence_info.xlsx)
- Raw assay data
- Processed assay data (figures 2 and 3)
- Fasta file of sequences tested in study
    - We should also deposit our sequences to addgene
- Identity with characterised (all EC)
- Anything else?

Author contributions

Xyz

Conflicts of interest

The authors declare no conflict of interest.

References

Alley, Ethan C., Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M. Church. 2019. “Unified Rational Protein Engineering with Sequence-Based Deep Representation Learning.” *Nature Methods* 16 (12): 1315–22.

Bastard, Karine, Adam Alexander Thil Smith, Carine Vergne-Vaxelaire, Alain Perret, Anne Zaparucha, Raquel De Melo-Minardi, Aline Mariage, et al. 2014. “Revealing the Hidden Functional Diversity of an Enzyme Family.” *Nature Chemical Biology* 10 (1): 42–49.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. “BLAST+: Architecture and Applications.” *BMC Bioinformatics* 10 (December): 421.

Chang, Yi-Chien, Zhenjun Hu, John Rachlin, Brian P. Anton, Simon Kasif, Richard J. Roberts, and Martin Steffen. 2016. “COMBREX-DB: An Experiment Centered Database of Protein Function: Knowledge, Predictions and Knowledge Gaps.” *Nucleic Acids Research* 44 (D1): D330–35.

Choi, Lee Jun, and Nur’aini Abdul Rashid. 2008. “Adapting Normalized Google Similarity in Protein Sequence Comparison.” *2008 International Symposium on Information Technology*. https://doi.org/10.1109/itsim.2008.4631601.

Cock, P. J. A., T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, et al. 2009. “Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btp163.

Danchin, Antoine, Christos Ouzounis, Taku Tokuyasu, and Jean-Daniel Zucker. 2018. “No Wisdom in the Crowd: Genome Annotation in the Era of Big Data - Current Status and Future Prospects.” *Microbial Biotechnology* 11 (4): 588–605.

Dellero, Younès, Caroline Mauve, Edouard Boex-Fontvieille, Valérie Flesch, Mathieu Jossier, Guillaume Tcherkez, and Michael Hodges. 2015. “Experimental Evidence for a Hydride Transfer Mechanism in Plant Glycolate Oxidase Catalysis.” *The Journal of Biological Chemistry* 290 (3): 1689–98.

Edgar, Robert C. 2004. “MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput.” *Nucleic Acids Research* 32 (5): 1792–97.

El-Gebali, Sara, Jaina Mistry, Alex Bateman, Sean R. Eddy, Aurélien Luciani, Simon C. Potter, Matloob Qureshi, et al. 2019. “The Pfam Protein Families Database in 2019.” *Nucleic Acids Research* 47 (D1): D427–32.

Enright, A. J., S. Van Dongen, and C. A. Ouzounis. 2002. “An Efficient Algorithm for Large-Scale Detection of Protein Families.” *Nucleic Acids Research* 30 (7): 1575–84.

Erb, Tobias J. 2019. “Back to the Future: Why We Need Enzymology to Build a Synthetic Metabolism of the Future.” *Beilstein Journal of Organic Chemistry* 15 (February): 551–57.

Esser, Christian, Anke Kuhn, Georg Groth, Martin J. Lercher, and Veronica G. Maurino. 2014. “Plant and Animal Glycolate Oxidases Have a Common Eukaryotic Ancestor and Convergently Duplicated to Evolve Long-Chain 2-Hydroxy Acid Oxidases.” *Molecular Biology and Evolution* 31 (5): 1089–1101.

Furnham, Nicholas, John S. Garavelli, Rolf Apweiler, and Janet M. Thornton. 2009. “Missing in Action: Enzyme Functional Annotations in Biological Databases.” *Nature Chemical Biology* 5 (8): 521–25.

Gerlt, John A. 2018. “The Need for Manuscripts To Include Database Identifiers for Proteins.”

Biochemistry 57 (29): 4239–40.
Gilks, Walter R., Benjamin Audit, Daniela de Angelis, Sophia Tsoka, and Christos A. Ouzounis. 2005. “Percolation of Annotation Errors through Hierarchically Structured Protein Sequence Databases.” *Mathematical Biosciences* 193 (2): 223–34.
Griesemer, Marc, Jeffrey A. Kimbrel, Carol E. Zhou, Ali Navid, and Patrik D’haeseleer. 2018. “Combining Multiple Functional Annotation Tools Increases Coverage of Metabolic Annotation.” *BMC Genomics* 19 (1): 948.
Guo, Xiaoting, Manman Zhang, Menghao Cao, Wen Zhang, Zhaoqi Kang, Ping Xu, Cuiqing Ma, and Chao Gao. 2018. “D-2-Hydroxyglutarate Dehydrogenase Plays a Dual Role in L-Serine Biosynthesis and D-Malate Utilization in the Bacterium Pseudomonas Stutzeri.” *The Journal of Biological Chemistry* 293 (40): 15513–23.
Hackenberg, Claudia, Ramona Kern, Jan Hüge, Lucas J. Stal, Yoshinori Tsuji, Joachim Kopka, Yoshihiro Shiraiwa, Hermann Bauwe, and Martin Hagemann. 2011. “Cyanobacterial Lactate Oxidases Serve as Essential Partners in N2 Fixation and Evolved into Photorespiratory Glycolate Oxidases in Plants.” *The Plant Cell* 23 (8): 2978–90.
Helbert, William, Laurent Poulet, Sophie Drouillard, Sophie Mathieu, Mélanie Loiodice, Marie Couturier, Vincent Lombard, et al. 2019. “Discovery of Novel Carbohydrate-Active Enzymes through the Rational Exploration of the Protein Sequences Space.” *Proceedings of the National Academy of Sciences of the United States of America* 116 (13): 6063–68.
Huang, Hua, Chetanya Pandya, Chunliang Liu, Nawar F. Al-Obaidi, Min Wang, Li Zheng, Sarah Toews Keating, et al. 2015. “Panoramic View of a Superfamily of Phosphatases through Substrate Profiling.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (16): E1974–83.
Jeske, Lisa, Sandra Placzek, Ida Schomburg, Antje Chang, and Dietmar Schomburg. 2019. “BRENDA in 2019: A European ELIXIR Core Data Resource.” *Nucleic Acids Research* 47 (D1): D542–49.
Jones, Craig E., Alfred L. Brown, and Ute Baumann. 2007. “Estimating the Annotation Error Rate of Curated GO Database Sequence Annotations.” *BMC Bioinformatics* 8 (May): 170.
Jones, J. M., J. C. Morrell, and S. J. Gould. 2000. “Identification and Characterization of HAOX1, HAOX2, and HAOX3, Three Human Peroxisomal 2-Hydroxy Acid Oxidases.” *The Journal of Biological Chemistry* 275 (17): 12590–97.
Kean, Kelsey M., and P. Andrew Karplus. 2019. “Structure and Role for Active Site Lid of Lactate Monooxygenase from Mycobacterium Smegmatis : Structure of Lactate Monooxygenase.” *Protein Science: A Publication of the Protein Society* 28 (1): 135–49.
Knorr, Sebastian, Malte Sinn, Dmitry Galetskiy, Rhys M. Williams, Changhao Wang, Nicolai Müller, Olga Mayans, David Schleheck, and Jörg S. Hartig. 2018. “Widespread Bacterial Lysine Degradation Proceeding via Glutarate and L-2-Hydroxyglutarate.” *Nature Communications* 9 (1): 5071.
Koga, Yuichi, Kanako Konishi, Atsushi Kobayashi, Shigenori Kanaya, and Kazufumi Takano. 2019. “Anaerobic Glycerol-3-Phosphate Dehydrogenase Complex from Hyperthermophilic Archaeon Thermococcus Kodakarensis KOD1.” *Journal of Bioscience and Bioengineering* 127 (6): 679–85.
Li, W., and A. Godzik. 2006. “Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btl158.
Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. “Scikit-Learn: Machine Learning in Python.” *Journal of Machine Learning Research: JMLR* 12 (85): 2825−2830.

Radivojac, Predrag, Wyatt T. Clark, Tal Ronnen Oron, Alexandra M. Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, et al. 2013. "A Large-Scale Evaluation of Computational Protein Function Prediction." *Nature Methods* 10 (3): 221–27.

Rassaei, Liza, Wouter Olthuis, Seiya Tsujimura, Ernst J. R. Sudhölter, and Albert van den Berg. 2014. "Lactate Biosensors: Current Status and Outlook." *Analytical and Bioanalytical Chemistry* 406 (1): 123–37.

Repecka, Donatas, Vykintas Jauniskis, Laurynas Karpus, Elzbieta Rembeza, Jan Zrimec, Simona Poviloniene, Irmantas Rokaitis, et al. 2019. "Expanding Functional Protein Sequence Space Using Generative Adversarial Networks." *bioRxiv*. https://doi.org/10.1101/789719.

Richardson, Emily J., and Mick Watson. 2013. "The Automatic Annotation of Bacterial Genomes." *Briefings in Bioinformatics* 14 (1): 1–12.

Roberts, Richard J., Yi-Chien Chang, Zhenjun Hu, John N. Rachlin, Brian P. Anton, Revonda M. Pokrzywa, Han-Pil Choi, et al. 2011. "COMBREX: A Project to Accelerate the Functional Annotation of Prokaryotic Genomes." *Nucleic Acids Research* 39 (Database issue): D11–14.

Salzberg, Steven L. 2007. "Genome Re-Annotation: A Wiki Solution?" *Genome Biology* 8 (1): 102.

Schnoes, Alexandra M., Shoshana D. Brown, Igor Dodevski, and Patricia C. Babbitt. 2009. "Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies." *PLoS Computational Biology* 5 (12): e1000605.

Singh, Rajendra, Manoj Kumar, Anshumali Mittal, and Praveen Kumar Mehta. 2016. "Microbial Enzymes: Industrial Progress in 21st Century." *3 Biotech* 6 (2): 174.

Sukumar, N., S. Liu, W. Li, F. S. Mathews, B. Mitra, and P. Kandavelu. 2018. "Structure of the Monotopic Membrane Protein (S)-Mandelate Dehydrogenase at 2.2 Å Resolution." *Biochimie* 154 (November): 45–54.

Swainston, Neil, Antonio Baici, Barbara M. Bakker, Athel Cornish-Bowden, Paul F. Fitzpatrick, Peter Halling, Thomas S. Leyh, et al. 2018. "STRENDA DB: Enabling the Validation and Sharing of Enzyme Kinetics Data." *The FEBS Journal* 285 (12): 2193–2204.

Tipton, Keith F., Richard N. Armstrong, Barbara M. Bakker, Amos Bairoch, Athel Cornish-Bowden, Peter J. Halling, Jan-Hendrik Hofmeyr, et al. 2014. "Standards for Reporting Enzyme Data: The STRENDA Consortium: What It Aims to Do and Why It Should Be Helpful." *Perspectives in Science* 1 (1): 131–37.

Tsiafoulis, Constantinos G., Mamas I. Prodromidis, and Miltiades I. Karayannis. 2002. "Development of Amperometric Biosensors for the Determination of Glycolic Acid in Real Samples." *Analytical Chemistry* 74 (1): 132–39.

Umena, Yasufumi, Kazuko Yorita, Takeshi Matsuoka, Akiko Kita, Kiyoshi Fukui, and Yukio Morimoto. 2006. "The Crystal Structure of L-Lactate Oxidase from Aerococcus Viridans at 2.1A Resolution Reveals the Mechanism of Strict Substrate Recognition." *Biochemical and Biophysical Research Communications* 350 (2): 249–56.

UniProt Consortium. 2019. "UniProt: A Worldwide Hub of Protein Knowledge." *Nucleic Acids Research* 47 (D1): D506–15.

Vanacek, Pavel, Eva Sebestova, Petra Babkova, Sarka Bidmanova, Lukas Daniel, Pavel Dvorak, Veronika Stepankova, et al. 2018. "Exploration of Enzyme Diversity by Integrating Bioinformatics with Expression Analysis and Biochemical Characterization." *ACS Catalysis* 8 (3): 2402–12.

Weghoff, Marie Charlotte, Johannes Bertsch, and Volker Müller. 2015. "A Novel Mode of Lactate Metabolism in Strictly Anaerobic Bacteria." *Environmental Microbiology* 17 (3):

670–77.
Xia, Z. X., and F. S. Mathews. 1990. “Molecular Structure of Flavocytochrome b2 at 2.4 A Resolution.” *Journal of Molecular Biology* 212 (4): 837–63.
Zallot, Rémi, Nils Oberg, and John A. Gerlt. 2019. “The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways.” *Biochemistry* 58 (41): 4169–82.
Zielezinski, Andrzej, Susana Vinga, Jonas Almeida, and Wojciech M. Karlowski. 2017. “Alignment-Free Sequence Comparison: Benefits, Applications, and Tools.” *Genome Biology* 18 (1): 186.