



## REVIEW ARTICLE

# Identifying pseudoenzymes using functional annotation: pitfalls of common practice

Antonio J. M. Ribeiro , Jonathan D. Tyzack, Neera Borkakoti and Janet M. Thornton 

European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, UK

**Keywords**catalytic residue; database; enzyme;  
enzyme evolution; pseudoenzyme**Correspondence**A. J. M. Ribeiro, European Bioinformatics  
Institute, Wellcome Trust Genome Campus,  
Hinxton, Cambridge CB10 1SD, UK  
Tel: +44 1223 494397  
E-mail: ribeiro@ebi.ac.uk

Correction added on 12 August, after first  
online publication: The title was changed  
from "Identifying pseudoenzymes using  
functional annotation. How loss of function  
correlates with mutations in the catalytic  
site" to "Identifying pseudoenzymes using  
functional annotation: pitfalls of common  
practice."

(Received 21 August 2019, accepted 14  
November 2019)

doi:10.1111/febs.15142

Pseudoenzymes are proteins that are evolutionary related to enzymes but lack relevant catalytic activity. They are usually evolved from enzymatic ancestors that have lost their catalytic activities. The loss of catalytic function is one extreme amongst the other evolutionary changes that can occur to enzymes, like the changing of substrate specificity or the reaction catalysed. However, the loss of catalytic function events remain poorly characterised, except for some notable examples, like the pseudokinases. In this review, we aim to analyse current knowledge related to pseudoenzymes across a large number of enzymes families. This aims to be a review of the data available in biological databases, rather than a more traditional literature review. In particular, we use UniProtKB as the source for functional annotation and M-CSA (Mechanism and Catalytic Site Atlas) for information on the catalytic residues of enzymes. We show that explicit annotation of lack of activity is not exhaustive in UniProtKB and that a protocol using lack of catalytic annotation as an indication for lack of function can be an adequate alternative, after some corrections. After identifying pseudoenzymes related to enzymes in M-CSA, we were able to comment on their prevalence across enzyme families, and on the correlation between lack of catalytic function and the mutation of catalytic residues. These analyses challenge two common ideas in the emerging literature: that pseudoenzymes are ubiquitous across enzyme families and that mutations in the catalytic residues of enzyme homologues are always a good indication of lack of activity.

## Introduction

Pseudoenzymes are noncatalytic proteins that share a common ancestor with enzymes. Lysozyme and  $\alpha$ -lactalbumin where the first related enzyme/nonenzyme pair of proteins to be identified, through sequence homology, more than 50 years ago [1]. At the time,  $\alpha$ -lactalbumin was wrongly predicted to be an enzyme, an early mistake which was indicative of the difficulties to come in the identification and characterisation of pseudoenzymes. As discussed in this data review, even with more sophisticated tools and large amounts of

sequence, structural and functional data, correctly identifying pseudoenzymes is still a challenge.

Pseudoenzymes in some enzyme families, of which the pseudokinases are the most notable example, are well characterised and provide a framework for researchers to understand how pseudoenzymes evolved, and what are their current roles in the cell. Although inactive kinase domains had been experimentally identified previously [2,3], it was only after the analysis of the kinase complement of the human genome that the extent of their abundance was revealed. Out of a total

**Abbreviation**

M-CSA, Mechanism and Catalytic Site Atlas.

of 518 sequences, 50 lack at least one catalytic residue [4]. Catalytic deficient kinases have important roles in the cell, including the allosteric regulation of other kinases, acting as scaffold for the assembly of signalling components or regulation of transcription [5,6]. Many of these noncatalytic roles are shared with their enzyme counterparts, which exhibit both functions [7]. Pseudokinases have been shown to be prevalent throughout the tree of life, and while on average they form about 10% of most species kinomes, that percentage can be as high as 50% [8]. Further examples of pseudoenzymes, other than kinases, reviewed elsewhere [9,10], perform similar but also different functions. Some pseudophosphatases compete with the conventional phosphatase for their substrate, while some pseudoproteases regulate protein localisation in the cell, for example. Generally, allosteric regulation of a related enzymatic counterpart seems to be the most common role across all pseudoenzyme families [10]. Still another group of pseudoenzymes, those that have evolved from moonlighting enzymes, perform roles completely unrelated to their enzymatic counterpart [11]. Moonlighting proteins perform two or more independent functions, typically in different cells or tissues, which allows for differential expression and regulation of the different functions while keeping only one gene [12]. The fact that many pseudoenzymes share their noncatalytic role with their enzyme counterparts suggests an evolutionary path for these examples whereby a common ancestor had both roles and, after duplication, one of the copies eventually loses the catalytic function. These events support a subfunctionalisation evolutionary model, where each postduplication gene specialises in one of the ancestor's functions [13,14], instead of the classical duplication model where new functionalisation comes after gene duplication [15].

Identifying pseudoenzymes is challenging on different levels. Experimentally, to validate the lack of catalytic function, inferred pseudoenzymes should be screened against a panel of biologically relevant substrates, to make sure the protein does not catalyse the original or any other catalytic function (see ref. [16], for example). However, the most common type of experimental evidence comes from testing the activity for a single reaction, that of the most similar enzyme to the proposed pseudoenzyme. From a computational point of view, pseudoenzymes are typically identified by finding homologue sequences of enzymes that lack one or more catalytic residues [8,17,18]. Although this methodology is adequate when the effect of mutations in similar enzymes is known, we think the blind application of this rule to any enzyme family is likely to overestimate the number of pseudoenzymes, since a mutation of a

catalytic residue does not always imply loss of catalytic function. These mutations can be accommodated by the enzyme (if a similar residue that can perform the same role is present in the active site, for example), or may give origin to a new enzyme that is able to catalyse a different reaction. Conversely, the methodology will also fail to identify those pseudoenzymes with conserved catalytic sites but lack of catalytic ability, since lack of function can be caused by other mechanisms, like the obstruction of the binding site, loss of cofactor binding or loss of allosteric regulatory potential [19,20]. We discuss these types of errors in our analysis below. Finally, using databases with functional annotation to identify pseudoenzymes is also not straightforward. Using the lack of enzymatic annotation as an indication for lack of activity would be a simple way to systematically categorise proteins as enzymes or pseudoenzymes, but the method is prone to classification errors (absence of evidence is not evidence of absence). Furthermore, many databases use overall sequence similarity to extend annotation to less well-known sequences, which will necessarily categorise pseudoenzymes as enzymes. The explicit annotation of lack of function (due to scarce experimental data) is still limited.

We discuss each of these points in more detail in the rest of the review. We start by exploring how annotation from UniProtKB [21] can be used to identify enzymes and nonenzymes. Secondly, we find potential pseudoenzymes by identifying which of these nonenzymes are also homologous to enzymes in M-CSA [22], a database of enzyme mechanisms and catalytic sites. This set of pseudoenzymes is then analysed to understand their evolutionary history and assess how mutations in the catalytic residues, which are annotated in M-CSA, are associated with the loss of catalytic function.

## Using UniProt annotation and M-CSA homologues to identify pseudoenzymes

UniProtKB is a comprehensive database for protein sequences and related annotations [21]. In this section, we look at which types of annotations present in UniProtKB can be used to discriminate between enzymes and nonenzymes. In particular, we focus on the manually reviewed subset of UniProtKB, called Swiss-Prot. Available annotations relating to the lack of enzymatic activity can be broadly divided into implicit and explicit annotations. Implicit annotation refers to the use of lack of enzymatic annotation as an indication for lack of activity. For example, using the EC annotation [23], one can label all proteins annotated with an EC

number as enzymes, and proteins without EC annotation as nonenzymes. UniProtKB Keywords (KW) and GO annotation [24,25] can be used in the same manner. In general, all proteins in Swiss-Prot include these three types of annotation (EC, KW, GO), so they are a good starting point to infer lack of activity across the entire Swiss-Prot. For a small portion of proteins, UniProtKB curators explicitly annotate their lack of activity. These annotations, however, are not consistently applied and most are not standardised, but are rather captured as free text. Even with these limitations, these explicit annotations, especially the subset of those directly supported by experimental evidence, are the best available indications of lack of catalytic function. In the next two sections, we discuss how both implicit and explicit annotations can be used to identify nonenzymes in Swiss-Prot and among M-CSA homologues. The flowchart in Fig. 1 describes the steps taken to create the datasets discussed in the following sections.

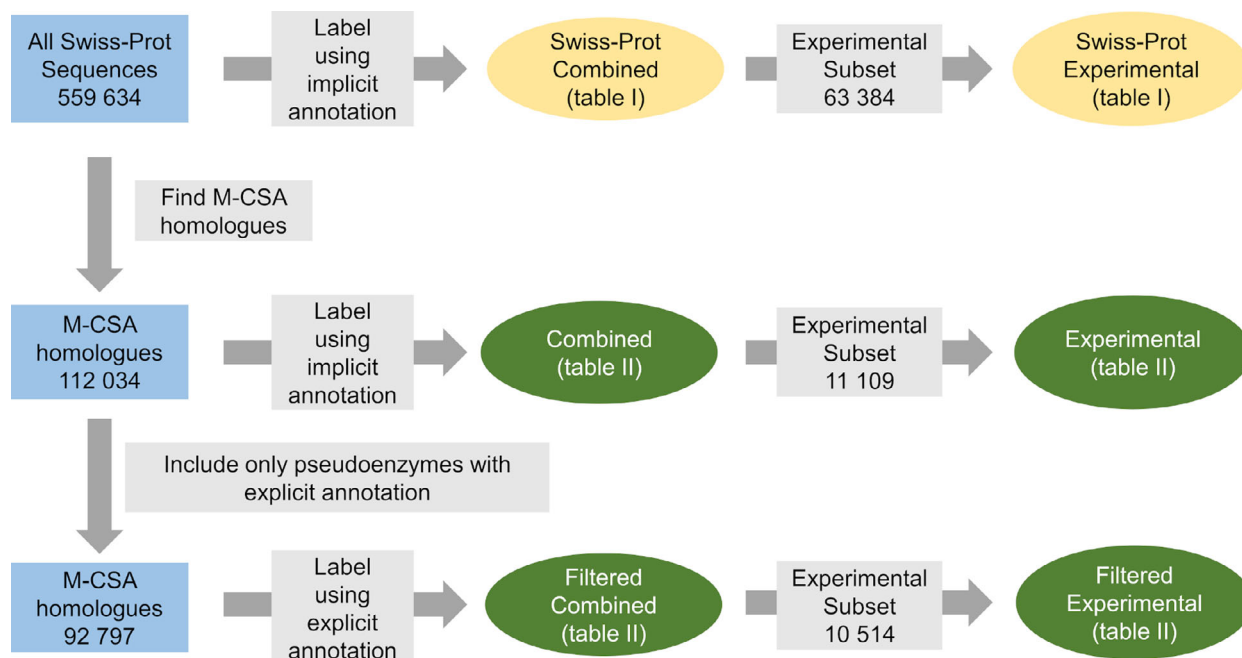
### Inferring lack of activity from implicit annotation in Swiss-Prot

We have looked at how the three types of enzymatic annotation (EC, KW & GO) described above can be used to infer lack of activity across the entire Swiss-Prot. Table 1 shows the numbers of enzymes and nonenzymes identified according to each of these

methods, and one last method that combines annotations. A brief description of each type of annotation is given in the following paragraphs.

The EC aims to classify all enzyme reactions according to a four-level hierarchical system [23]. Enzymes in UniProtKB are assigned at least one EC number. Incomplete EC numbers (where not all the hierarchical levels are defined) can be used when there is not a more specific EC number to be given or the exact substrates of the enzyme are not known. In the first row of Table 1, we identify proteins as enzymes if they are annotated with at least one EC number or as nonenzymes otherwise. According to this rule, 47.6% of the sequences in Swiss-Prot are enzymes, and the remaining 52.4% are nonenzymes.

UniProtKB keywords are a hierarchical controlled vocabulary developed and maintained by UniProtKB. Keyword assignment is mostly a manual annotation process informed by the literature and also by sequence similarity. For the purposes of this analysis, we categorised as enzymes all the entries with one of the following catalytic keywords (including their child terms) under the 'Molecular Function' category: Oxidoreductase, Transferase, Hydrolase, Lyase, Isomerase, Ligase, Translocase, Motor Protein, ATP Synthesis and Electron Transport. Conversely, we classify proteins that do not have these keywords as nonenzymes. Note that the first 7 keywords in the list correspond to the main 7 EC classes. The last three keywords are



**Fig. 1.** The overall steps taken to create the datasets discussed in this paper.

**Table 1.** Number of enzymes and nonenzymes in Swiss-Prot according to different types of UniProtKB annotation.

Swiss-Prot	Enzymes		Nonenzymes	
All proteins				
EC	266 207	47.6%	293 427	52.4%
KW	281 483	50.3%	278 151	49.7%
GO	301 489	53.9%	258 145	46.1%
Swiss-Prot combined <sup>a</sup>	282 428	50.5%	277 206	49.5%
Total	559 634			
Experimental only				
Swiss-Prot experimental <sup>a</sup>	29 731	46.9%	33 653	53.1%
Total	63 384			

<sup>a</sup> In the Swiss-Prot combined dataset, proteins are considered enzymes if they are associated with either a catalytic KW or a catalytic experimental GO annotation. See main text for details. In the Swiss-Prot experimental dataset, the same rule is used but applied only to the sequences with functional annotation supported by experimental evidence.

associated with some proteins that are enzymes, but where the catalytic function is secondary to the main function, and probably for this reason, they are not consistently annotated as enzymes in UniProtKB. According to these rules, 50.3% of Swiss-Prot sequences are enzymes, 2.7% more than derived from EC annotation.

Proteins in UniProtKB are also given GO terms [24,25], aggregated from a variety of sources. For the values presented in Table 1, we categorised as enzymes the proteins annotated with a catalytic GO term (GO:0003824 – ‘catalytic activity’, and its children terms) plus an ‘enables’ qualifier. The attribution of GO terms in UniProtKB can be made by manual curation, using the literature, or be based on automatic methods, such as extending annotation to sequence homologues. Since, by definition, pseudoenzymes are homologous to enzymes, automatic annotation based on the overall sequence homology wrongly categorises pseudoenzymes as ‘catalytic’. This is the reason for the higher, but incorrect, number of enzymes according to GO annotation (53.9% of Swiss-Prot). However, the subset of GO associations directly supported by experimental evidence correctly identifies enzymes and pseudoenzymes. These experimental annotations can be filtered by selecting only ‘protein–GO term’ associations labelled with an ‘experimental evidence’ ECO code (ECO:0000006) [26] or its children terms. We use experimental GO annotations in the ‘combined’ sets.

After considering the advantages and limitations of the EC, KW and GO sets of enzymes and nonenzymes, we devised a simple rule to identify all the sequences in Swiss-Prot as either enzyme or

nonenzyme, for the purpose of this analysis. All the proteins with either a catalytic KW annotation or an experimentally supported catalytic GO annotation are considered enzymes. The remaining proteins are classified as nonenzymes. Since all EC annotations are transferred to KW annotation (all proteins with an EC number are given a catalytic KW), the combined rule does not need to consider EC annotation explicitly. GO annotations that are not experimentally supported are ignored. According to this ‘combined’ rule, 50.5% of the proteins in Swiss-Prot are enzymes and 49.5% are nonenzymes.

Finally, we consider another subset of Swiss-Prot, composed only of proteins for which there is functional experimental evidence. These are proteins where either the ‘function’, ‘misc’ or ‘caution’ fields in the UniProtKB entry or the catalytic GO terms are associated with the ECO code denoting experimental evidence. Only about 11% of Swiss-Prot, or 63 440 sequences, contain experimental evidence related to the function of the protein. Of these, 46.9% are enzymes and 53.1% are nonenzymes.

## Identification of potential pseudoenzymes among M-CSA homologues

Up to this point, we have been discussing the categorisation of Swiss-Prot proteins as either enzymes or nonenzymes. In order to identify pseudoenzymes among these, we must identify which nonenzymes are also related to enzymes. Since one of the purposes of this analysis is to study the correlation between mutation of the catalytic residues and loss of function, we use the enzymes in M-CSA database, for which we have annotations of the catalytic residues, as our enzyme reference set. To find the homologues, we have searched Swiss-Prot for proteins similar to the enzymes in M-CSA, using phmmer with a cut-off *e*-value of  $10^{-10}$ . This cut-off value is relatively stringent and will not find very distant homologues. However, it makes the alignment of the catalytic residues in the homologue sequences and the construction of the phylogenetic trees less speculative. While this is essential for our analysis, this choice will impact the number of pseudoenzymes we identify per enzyme family. A deeper search, by finding more distant homologues, would probably find a greater number of pseudoenzymes.

After a preliminary analysis, we noticed that most proteins identified as pseudoenzymes in this manner belonged to a small number of families for which annotation in UniProtKB was erroneously missing enzymatic annotation, and so, these were not real

pseudoenzymes. For example, proteins belonging to the 'Heat shock protein 70 family' have ATPase activity, which is required for the correct function of the protein, but annotation for that activity is missing for many sequences in the family. Prompted by this, we manually checked all the enzyme families with pseudoenzymes (using the phylogenetic trees described below) to identify similar problems. We identified 17 families, defined by their PFAM domains [27], which were excluded from further analysis: Heat shock protein 70 family, Chaperonin Cpn60/TCP-1 family, Transcription factor, GTP-binding domain, Phosphopantetheine binding ACP domain, Glycine radical domain, ATPase AAA-type core, Small GTPase superfamily ARF/SAR type, Guanine nucleotide-binding protein (G-protein), Nitrogenase/oxidoreductase component 1, gag p10 and gag p24, DsbC protein family, Radical SAM, Strictosidine synthase, Thioredoxin, Kinesin motor domain 6 and ATP synthase alpha/beta subunits.

Using the clean dataset, and as shown in Table 2, 112 034 Swiss-Prot proteins are also M-CSA homologues (20.0%). The percentage of enzymes in Swiss-Prot covered by the M-CSA dataset is higher, closer to 40%. The latter number is a good estimation of the M-CSA coverage of Swiss-Prot.

We now consider the UniProt annotations for the 112 034 M-CSA homologues we found. The number of pseudoenzymes identified based on lack of the

annotation related to catalytic activity in UniProtKB varies significantly for the three types of annotations (EC, KW and GO). Only 4520 (4.0%) of the enzyme homologues lack EC annotation; 1296 (1.2%) lack a catalytic KW; and a mere 619 (0.6%) are not annotated with a catalytic GO term. The caveats described in the previous section for each of these annotations also apply here, as well as the 'combined' rule, which combines KW and experimental GO annotation, and estimates that 1277 (1.1%) of the M-CSA homologues in Swiss-Prot are pseudoenzymes. With respect to proteins with functional experimental annotation, the estimation for the number of pseudoenzymes in this dataset is more than double, at 3.2% or 361 of the 11 109 identified homologues, which means that pseudoenzymes in Swiss-Prot are more commonly annotated with experimental data compared to enzymes, relative to their abundance.

## Identifying pseudoenzymes using explicit lack of function annotation

By using the lack of annotation to infer lack of activity, as described in the previous sections, it is possible to label all the Swiss-Prot proteins as enzymes or nonenzymes, and most M-CSA homologues as enzymes or pseudoenzymes. The obvious limitation of this methodology is that the lack of annotation is not equivalent to lack of activity, and so, it is likely that some enzymes have been wrongly identified as pseudoenzymes. For this reason, we decided to create two extra subsets of enzymes/pseudoenzymes with a stricter set of rules (listed below), which include only pseudoenzymes for which the lack of activity is explicitly annotated. The starting point for the creation of these two sets ('combined filtered' and 'experimental filtered') are the 'combined' and 'experimental' rules described before. In the filtered sets, only sequences that obey at least one of the following rules are considered pseudoenzymes:

- 1 The UniProtKB name of the sequence starts with 'inactive', 'probable inactive', 'putative inactive' or 'probably inactive'. This is a nonstandardised and nonexhaustive flag that UniProtKB curators attribute specifically to pseudoenzymes.
- 2 The protein is annotated with at least a catalytic GO term with a 'NOT – enables' qualifier. In our opinion, this is the most robust way of capturing lack of activity information, since it is easy to filter, it is associated with an ECO code and allows for annotation about multiple enzyme activities.

**Table 2.** Number of enzymes and nonenzymes homologous to 964 M-CSA entries according to different types of UniProtKB annotation. Nonenzymes in this table are also pseudoenzymes.

M-CSA homologues	Enzymes		Pseudoenzymes	
Inferred lack of activity				
All proteins				
EC	107 514	96.0%	4520	4.0%
KW	110 738	98.8%	1296	1.2%
GO	111 415	99.4%	619	0.6%
Combined <sup>a</sup>	110 757	98.9%	1277	1.1%
Total	112 034			
Experimental only				
Experimental <sup>a</sup>	10 748	96.8%	361	3.2%
Total	11 109			
Annotated lack of activity				
All proteins				
Filtered combined <sup>a</sup>	92 288	99.5%	503	0.5%
Total	92 791			
Experimental only				
Filtered experimental <sup>a</sup>	10 328	98.2%	186	1.8%
Total	10 514			

<sup>a</sup> These four sets are the ones considered in the subsequent analyses.



3 The protein does not have a catalytic KW and the 'function', 'caution' or 'misc' fields in the UniProt entry explicitly mention that the protein lacks or probably lacks enzyme activity. These annotations are first filtered automatically based on expressions like 'lacks activity', and then manually checked.

We also apply stricter filters to both enzyme and nonenzyme sequences in these sets. In particular, we exclude sequences that lack any text in the 'function', 'caution', 'misc' fields and noncharacterised proteins, whose UniProt names start with 'Uncharacterized'. Lastly, we remove from the dataset any sequences for which we cannot align all the reference catalytic residues. This filter is particularly useful to eliminate from the analysis proteins with active sites that are composed of amino acids from different domains, and where the lack of activity may be attributed to the lack of part of the protein rather than specific mutations in the active site. After these stricter rules are applied, the number of sequences in the 'Combined Filtered' dataset is 92 791, of which 503 (0.5%) are pseudoenzymes. For the 10 514 sequences with experimental evidence, 186 (1.8%) are explicitly annotated as lacking catalytic activity.

It is observed that the number of pseudoenzymes identified in this manner is quite low. We would expect the real number of pseudoenzymes in these enzyme families to be higher, closer to the nonfiltered datasets. Together with the stringent filters, the strict homology search performed in this analysis is probably driving the number of pseudoenzymes down. It is likely that deeper homology searches, by finding more distantly related sequences, will retrieve more examples of pseudoenzymes, since distant relatives are more probable to have changed their functions. Nonetheless, the current method is a considerable improvement to using solely the lack of one type of annotation, which can grossly overestimate the number of pseudoenzymes. For example, the number of proteins missing EC classification is almost 4 times larger than the number of pseudoenzymes detected according to the 'combined' rule (Table 2).

### Visualising pseudoenzyme evolution using sequence trees

In the previous sections, we have abstracted four enzyme/nonenzyme sets, built by filtering Swiss-Prot sequences according to different sets of rules, which were labelled: Combined, Experimental, Filtered Combined, and Filtered Experimental. For each one of these protein sets and for each reference enzyme in

M-CSA, we have created an annotated evolutionary tree that shows the evolutionary history of these enzymes/nonenzymes and the conservation of the catalytic residues, where these are defined by the reference M-CSA enzyme.

To create each tree, we started by searching the reference sequence of that enzyme in M-CSA against a search dataset, using phmmer [28] with a cut-off e-value of  $1 \times 10^{-10}$ . The search dataset includes all the proteins in Swiss-Prot and proteins belonging to the reference proteomes of UniProt [21] that have at least one protein in Swiss-Prot (the filters discussed previously are not applied at this point). Although proteins from the reference proteomes are not included in the final trees or subsequent analysis, these add extra information that improves the alignments and the topology of the evolutionary trees. The sequences found by phmmer are aligned using the multiple sequence alignment tool MAFFT [29]. At this step, the number of proteins from reference proteomes in each alignment is capped at 5000, to avoid very large and computationally demanding alignments. Each multiple sequence alignment, after removal of poorly aligned regions with trimAl [30], was used to construct a maximum likelihood evolutionary tree using FASTTREE [31]. These sequence-based trees are then reconciled with the NCBI [32] species phylogenetic tree using Notung [33], where branches with weak sequence support (< 75%) are reordered to minimise gene duplication events. The trees were rooted following the same principle. Since the homology search is not deep enough, due to the use of a strict e-value cut-off and noniterative search algorithm, and because we have a cap on the total number of sequences included in the alignment, gene loss events, which would be erroneously predicted, are ignored. For each of the four sets considered, the rooted and reconciled trees are then pruned to include only the sequences that belong to that set. Finally, we annotate each tree with the catalytic residues of the M-CSA reference sequence and their respective position and conservation status in the homologous sequences. Each sequence in the tree is also annotated as enzyme/nonenzyme, and with a list of EC numbers [23] and PFAM domains [27]. We use a parsimonious algorithm to infer the position of the loss/gain of function events in the protein tree that tries to find the minimum number of events that explain the enzyme/nonenzyme distribution. The BIO-PHYLO [34,35] and ETE3 [36] toolkits were used to manipulate the sequence trees and to download the subset of the NCBI species tree relevant for each reconciliation. The annotated protein trees for every entry in M-CSA and the four protein sets can be

browsed and visualised at [www.ebi.ac.uk/thornton-srv/m-csa/tree](http://www.ebi.ac.uk/thornton-srv/m-csa/tree). We use the PhyD3 javascript viewer [37] to show the annotated trees in the website. In the next section, we show and comment on the annotated protein tree built for lysozyme (P00698, M-CSA:203).

### Case study – revisiting the Lysozyme/ $\alpha$ -Lactalbumin shared ancestry

Bovine  $\alpha$ -Lactalbumin and hen egg white Lysozyme are considered to be the first pseudoenzyme/enzyme pair to be identified, by sequence homology, more than 50 years ago [1]. In the original work, at least one of the matching catalytic residues of lysozyme was confirmed to be mutated in  $\alpha$ -lactalbumin (Glu35->His; Glu35 corresponds to Glu53 in Fig. 2, and the current protocol based on the multiple sequence alignment aligns this Glu with a Thr in P00711, rather than a neighbouring His), but the authors assumed the function of  $\alpha$ -lactalbumin to be enzymatic. The catalytic mechanism was unknown at the time, but since Glu and His can do some of the same catalytic roles, that is, both can act as a general acid–base, it was hypothesised that despite the mutation a catalytic function could still be performed. In fact,  $\alpha$ -lactalbumin is the regulatory subunit of the lactose synthase heterodimer (LS) and it does not perform any catalytic function itself. Here, we look at the lysozyme evolutionary tree, shown in Fig. 2 as an example for patterns and observations seen in other protein trees containing both enzymes and nonenzymes.

Lysozymes catalyse the hydrolysis of a glycosidic bond in the peptidoglycan of Gram-positive bacterial cell wall, causing the digestion of the wall and eventual bursting of the cell membrane. Chicken lysozyme C is annotated in M-CSA with six catalytic residues: Glu53, Asn64, Asp66, Ser68, Asp70 and Asn77. Two of these have a direct role in the reaction: Asp70 is the nucleophile and Glu53 acts as a general acid/base [38]. The remaining four are part of a hydrogen bond network that is important for substrate binding and catalysis, but they are not directly involved in the formation and cleavage of bonds [39,40]. All the nonenzymes in the lysozyme tree have at least one of the two main catalytic residues mutated, while most of the enzymes have both residues conserved. The only exceptions to this rule are the mouse, human and cattle sperm acrosome-associated protein 5, where Asp70 is mutated to Glu. The four catalytic residues not directly involved in the reaction are very well conserved in lysozyme C, although not universally, and there is no clear difference between nonenzymes and other lysozyme-like sequences. These results show a

good correlation between mutations in the catalytic residues and loss of catalytic function, but the correlation is not absolute and does not apply for every residue.

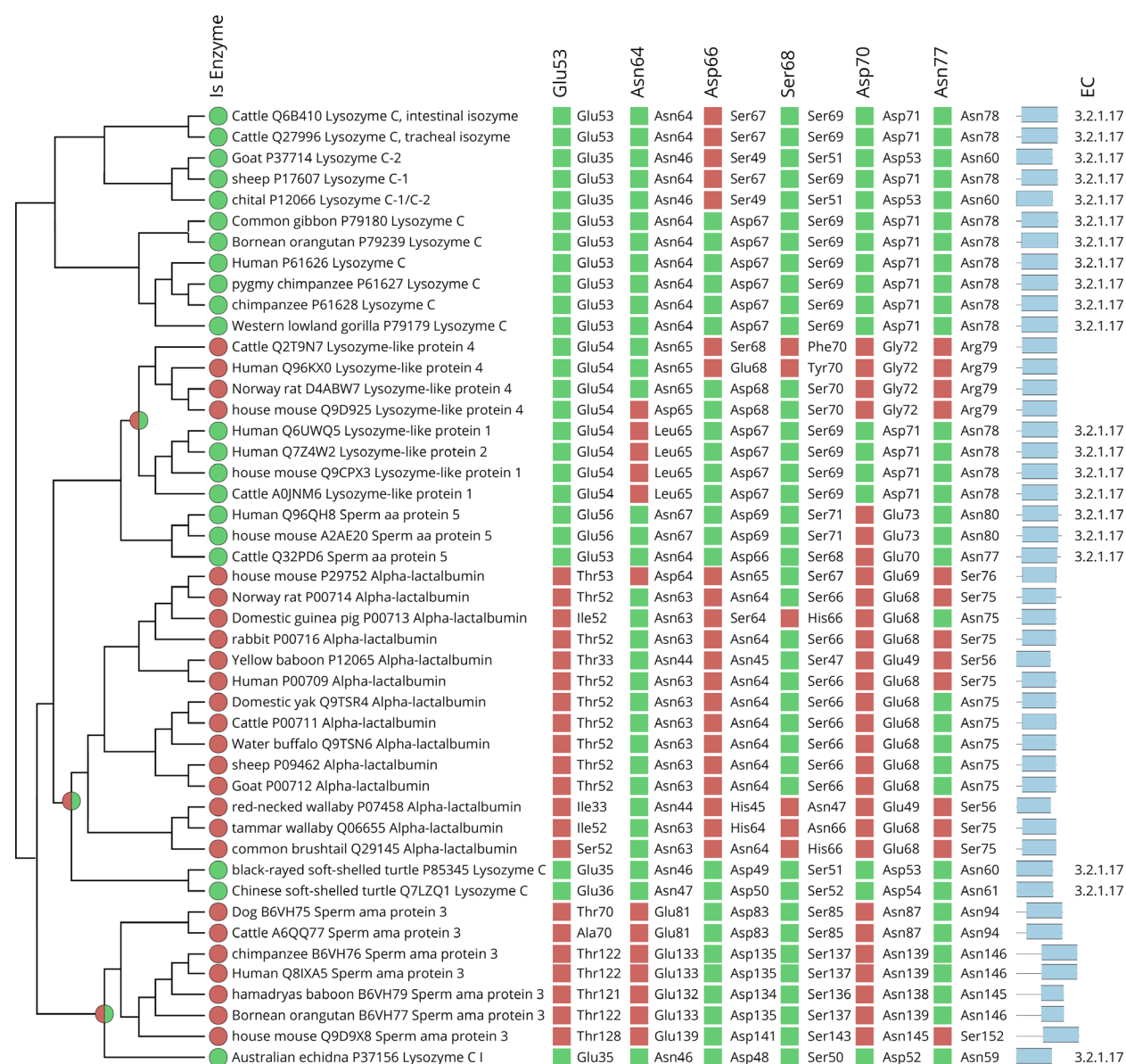
The topology of the tree provides extra insight. It suggests the existence of three loss-of-function events, indicated by the half green/half red circles superimposed on the tree. These events gave origin to three distinct groups of pseudoenzymes, which have different mutation profiles. In the first group (from top to bottom), both the equivalent residues of Asp70 and Asn77 are mutated, whilst the second group lost Glu53, Asp66 and Asp70, and the third group lost Glu53, Asn64 and Asp70. Together with the tree's topology, which was built from the entire sequence and species phylogeny, the mutation profiles of the catalytic residues provide further evidence that these groups of pseudoenzymes evolved separately and each group shares a separate common ancestor.

### Evolution and prevalence of pseudoenzymes across enzyme families and species

Most enzyme families in M-CSA do not contain any pseudoenzymes, as defined with the set of rules described previously (see Fig. 3). The number of families with pseudoenzymes also decreases as we apply stricter filters in annotation, from 210 (22.9%) in the 'combined' dataset to 87 (10.7%) in the 'filtered experimental' dataset. The difference in values highlights the sensitivity of this kind of analysis to the available annotation. Furthermore, this number is also dependent on the method used to find the enzyme homologues, as discussed previously, and on the M-CSA dataset, which is not exhaustive and includes homologous enzymes, as long as their mechanisms are different. We expect our conservative approach to be a lower bound to the number of pseudoenzymes present in the studied enzyme families.

We find that most families with both kinds of proteins have a small number of pseudoenzymes, <10%, and the number of families with more than 30% of nonenzymes is quite small, only 4 for the 'filtered experimental' dataset. In fact, the number of families where only one pseudoenzyme is identified represents 35% (combined set) to 45% (filtered experimental set) of the families with pseudoenzymes. The breakdown of the number of enzymes and pseudoenzymes in all M-CSA enzyme families is given at [www.ebi.ac.uk/thornton-srv/m-csa/tree](http://www.ebi.ac.uk/thornton-srv/m-csa/tree).

The phylogenetic trees inform us about how common are the loss- and gain-of-function events, that is,

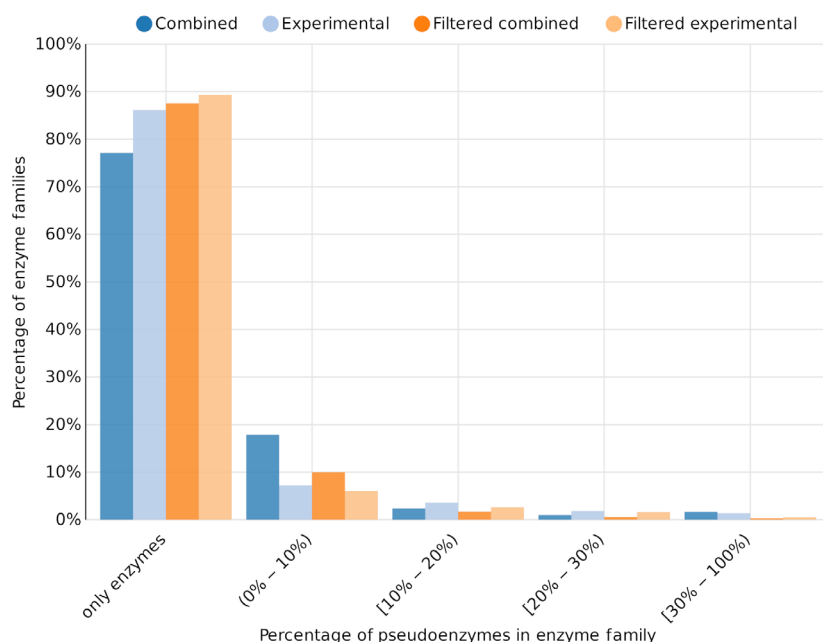


**Fig. 2.** Detail of the evolutionary protein tree built from the multisequence alignment of lysozyme Swiss-Prot homologues. Some sequences were removed from the tree for simplicity. The circle at the left of each species name indicates if the protein is an enzyme (green) or a nonenzyme (red). The half green/half red circles drawn on top of the tree represent bifurcations in function and possible points where the protein lost the catalytic activity. All the enzymes in this tree have the same EC number for lysozyme (EC 3.2.1.17). Squares indicate if the catalytic residue is conserved (green) or mutated (red). The blue rectangle represents the PFAM domain PF00062 (Glycoside hydrolase, family 22) to which all the proteins in the tree belong. The figure was created using the phylogenetic viewer PhyD3.

all the pseudoenzymes in the same tree can share the same ancestor or instead have independent origins. There is no clear correlation between the number of nonenzymes in a tree and the number of loss-of-function events. For some families, many recent (in evolutionary terms) but independent loss-of-function events are responsible for the large number of pseudoenzymes, while in others, the origins of several different

pseudoenzymes can be traced back to the same ancestor. Families with several independent loss-of-function events include serine proteases, protein kinases and phospholipases. Families where one ancient loss-of-function event gave origin to a large family of pseudoenzymes include DNA photolyases, from which the cryptochrome nonenzymes evolved, and the already discussed  $\alpha$ -lactalbumin, which evolved from lysozyme.





**Fig. 3.** The distribution of pseudoenzymes in M-CSA enzyme families. The first group of columns refers to the number of enzyme families with no pseudoenzymes. The percentages in the x-axis are the proportion of pseudoenzymes in each family. The total number of families for each set is 919, 822, 901 and 817, respectively. The different sets of proteins considered are the ones shown in Table 2.

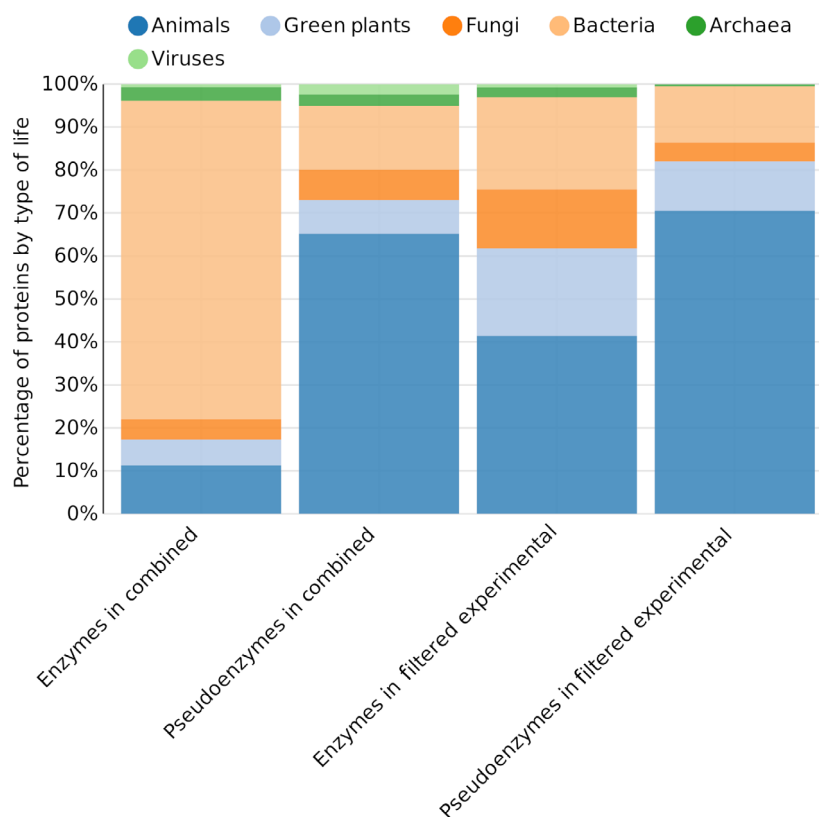
In some families, gain-/loss-of-function events are predicted to follow other older events. These cases are not straightforward to analyse because without reconstruction and testing of the ancient proteins, it is impossible to distinguish between a loss and regain of function or two losses of function positioned downstream. For this reason, in the previous discussion and the tree tables at the M-CSA website, we ignore all events in the subtrees after the first event. In general, events of gain of catalytic function are very rare in our dataset. In fact, we do not find any case where the original event is a gain of function. In some families (22 in the ‘combined’ set), the enzyme/nonenzyme common ancestor is at the root of the tree, and so, no prediction can be made about the activity of that ancestor, suggesting that gain-of-function events could have happened in these families at a very early evolutionary time.

Figure 4 shows how enzymes and pseudoenzymes are distributed across species for the ‘combined’ and ‘filtered experimental’ datasets. For both datasets, it is visible that pseudoenzymes are expanded in animals in comparison to other types of life. The low number of pseudoenzymes in bacteria is particularly striking, since most enzymes in the combined dataset belong to bacteria by a large margin. In the filtered experimental dataset, the difference is not so large, since the number of enzymes in bacteria is smaller to start with (which shows a bias of experimental evidence towards eukaryotes). These observations are in line with the distribution of kinases and pseudokinases across the tree of

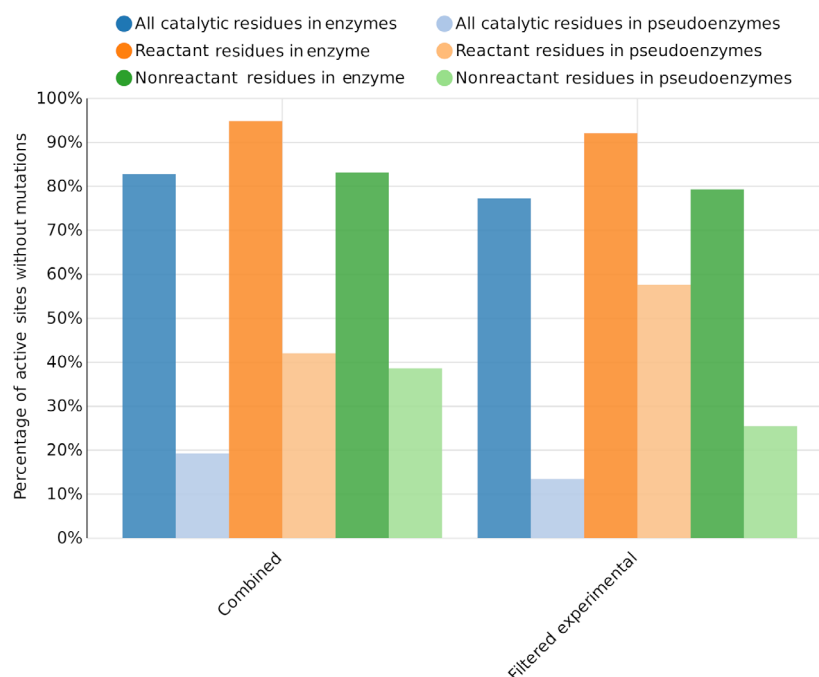
life [8]. The greater need for regulation in more complex organisms is probably driving the enrichment for pseudoenzymes in animal genomes.

### Mutation of catalytic residues and loss of function

A common way to identify potential pseudoenzymes relies on the identification of enzyme homologues where the ‘catalytic’ residues are mutated. However, examples of pseudoenzymes with conserved active sites have been described before [5,19], and mutation of a catalytic residue does not necessarily impair activity. In this section, we test the adequacy of this rule in our datasets. Figure 5 shows the conservation of active sites in enzymes and pseudoenzymes in the two datasets under study and for different groups of catalytic residues. The data show that when considering complete active sites (blue bars), the rule points in the right direction, that is, pseudoenzymes have more mutations in catalytic residues compared to enzymes, but it grossly overestimates the number of pseudoenzymes. For the stricter dataset, for example, more than 20% of the enzymes have at least one catalytic residue mutated (note that this set includes both enzymes that catalyse the same function, or a different function), so this rule would predict 20% of the enzymes in the dataset to be pseudoenzymes. Conversely, more than 10% of the pseudoenzymes have all the catalytic residues conserved, so these would be wrongly categorised as enzymes.



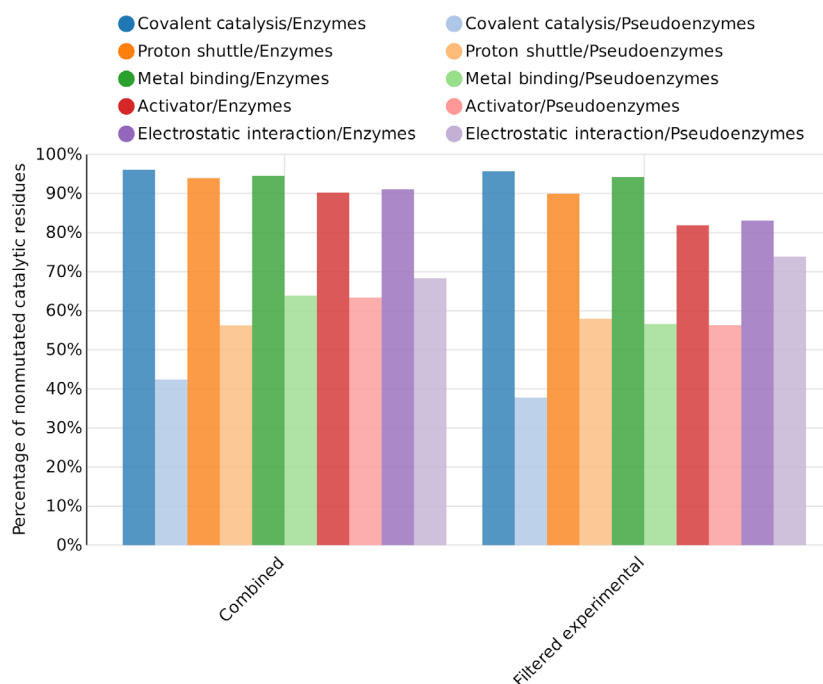
**Fig. 4.** The distribution of enzymes and pseudoenzymes in the in eukaryotes and other types of life. Eukaryotic cells other than animals, green plants and fungi are excluded from the plot, for simplicity.



**Fig. 5.** Conservation of the catalytic sites and catalytic residues in enzymes and pseudoenzymes. For the blue columns, each active site is flagged as conserved if all the catalytic residues are conserved and as not conserved if there is at least one mutation. For columns in orange and green, groups of residues are analysed separately. For example, for orange columns, active sites are flagged as conserved if all residues that have a reactant role are conserved. Catalytic role definitions follow the Enzyme Mechanism Ontology (EMO) [41]. The different sets of proteins considered are the ones shown in Table 2.

Based on these results, it is necessary to be cautious about using ‘mutation of catalytic residues’ as the only criterion to identify lack of activity. In our opinion, a rule like this is adequate when applied to characterised

enzyme families where specific mutations are known to impair function [8,18], but its use for the bulk identification of pseudoenzymes in poorly categorised enzyme families is likely to grossly overstate the number of



**Fig. 6.** Conservation of individual catalytic residues in enzymes and pseudoenzymes by the role they catalyse. Catalytic roles definitions follow the Enzyme Mechanism Ontology (EMO) [41]. The different sets of proteins considered are the ones shown in Table 2.

pseudoenzymes. Enzymes with mutated active sites may have compensatory mutations or be catalysing different reactions that do not require the original mutated residue.

When looking separately at groups of catalytic residues that perform reactant (directly involved in the formation/cleavage of bonds) or nonreactant roles, it is clear that reactant residues are more powerful at filtering out erroneously categorised pseudoenzymes. However, that comes with the cost of being able to identify fewer pseudoenzymes, since the fraction of pseudoenzymes with conserved reactant catalytic residues is also higher. (See the M-CSA website or previous publications [41] for a description of the Enzyme Mechanism Ontology used to define these roles.)

A more fine-grained approach compared to looking at the conservation at the active site level is to consider each catalytic residue separately. Figure 6 shows the conservation of catalytic residues in enzymes and pseudoenzymes categorised by their catalytic role. Some roles, like covalent catalysis (formation of covalent bonds between the residue and substrate), have better discriminatory power than others, like electrostatic stabilization, but no role provides a perfect rule. The results in this plot, together with the active site conservation results shown in Fig. 5 and the mutation profiles observed in the phylogenetic trees (see Fig. 2), suggest that better rules for the identification of pseudoenzymes among uncharacterised sequences are both possible and needed. Despite the requirements for

experimental confirmation, against a panel of substrates if possible, better bioinformatics tools based on existing annotation and reaction mechanistic information could make the identification of potential pseudoenzymes much more robust.

## Conclusion

In this data review, we have provided an analysis of the current knowledge about pseudoenzymes available in two biological databases. With respect to their identification, lack of activity of most of the nonenzymes can be inferred from the lack of enzymatic annotation, but this method has pitfalls, as shown by the lack of agreement among three types of annotation: EC, UniProt Keyword and GO. We have defined a general rule, based on KW annotation and experimental GO associations, that seems to eliminate most mistakes. Even so, we found some enzyme families which were not consistently annotated in UniProtKB, especially when the main function of most proteins in the family is not the enzymatic one.

For nonenzymes that are homologous to enzymes in M-CSA, we were able to define a stricter set of pseudoenzymes for which explicit annotation of lack of function is available. These validated pseudoenzymes represent less than 2% of all the proteins in the alignments of homologous sequences. Due to the very strict filters we applied to our data and homology search, this percentage of pseudoenzymes can be considered a

lower-bound estimate to the real number of pseudoenzymes. Deeper homology searches, in particular those based on structure, will find more pseudoenzymes [10,42]. On the other hand, we show that estimates based uniquely on the lack of catalytic annotation (using EC, for example) are likely to overestimate the number of pseudoenzymes.

We hope to have shown that the annotated phylogenetic trees containing both enzymes and related pseudoenzymes are a powerful way to describe their evolutionary history and the events that led to the formation of pseudoenzymes. The trees created for the M-CSA reference sequences indicate that most enzymatic families do not have validated pseudoenzymes, but that loss-of-function events are relatively common and can occur multiple times independently in the same enzyme family. We provide evidence that more sophisticated rules for discriminating enzymes and nonenzymes are needed, since a simple rule based on the mutation of any catalytic residue will wrongly categorise a large number of enzymes as pseudoenzymes (more than 20% in our stricter dataset) and miss the identification of some pseudoenzymes (more than 10% in the same dataset). We show that considering the role of the catalytic residue being mutated has better predictive power but is not a perfect solution. We consider that combining mechanistic information about the catalytic residues with the mutation profiles apparent from the phylogenetic trees could be a way forward to create better predictors for loss of function. We are working on a machine learning method that integrates these kinds of data to classify uncharacterised proteins similar to enzymes as either the same enzyme, an enzyme that catalyses a different reaction or a pseudoenzyme.

## Acknowledgements

The authors would like to thank Jean O'Driscoll for her contribution to the early stages of this analysis, and Rossana Zaru, Gemma Holliday and Roman Laszkowski for helpful discussions. The authors would like to thank EMBL for funding.

## Conflicts of interest

The authors declare no conflict of interest.

## Author contributions

All authors were involved in the conception and design of the work and revising the manuscript. AJMR developed the analysis tools and wrote the first version of

the manuscript. JMT was responsible for supervising the project.

## References

- 1 Brew K, Vanaman TC & Hill RL (1967) Comparison of the amino acid sequence of bovine  $\alpha$ -Lactalbumin and hens egg white lysozyme. *J Biol Chem* **242**, 3747–3748.
- 2 Gurniak CB & Berg LJ (1996) A new member of the Eph family of receptors that lacks protein tyrosine kinase activity. *Oncogene* **13**, 777–786.
- 3 Lim KL, Kolatkar PR, Ng KP, Ng CH & Pallen CJ (1998) Interconversion of the kinetic identities of the tandem catalytic domains of receptor-like protein-tyrosine phosphatase PTP $\alpha$  by two point mutations is synergistic and substrate-dependent. *J Biol Chem* **273**, 28986–28993.
- 4 Manning G (2002) The protein kinase complement of the human genome. *Science* **298**, 1912–1934.
- 5 Boudeau J, Miranda-Saavedra D, Barton GJ & Aless DR (2006) Emerging roles of pseudokinases. *Trends Cell Biol* **16**, 443–452.
- 6 Byrne DP, Foulkes DM & Evers PA (2017) Pseudokinases: update on their functions and evaluation as new drug targets. *Future Med Chem* **9**, 245–265.
- 7 Kung JE & Jura N (2016) Structural basis for the non-catalytic functions of protein kinases. *Structure* **24**, 7–24.
- 8 Kwon A, Scott S, Taujale R, Yeung W, Kochut KJ, Evers PA & Kannan N (2019) Tracing the origin and evolution of pseudokinases across the tree of life. *Sci Signal* **12**, eaav3810.
- 9 Murphy JM, Farhan H & Evers PA (2017) Bio-Zombie: the rise of pseudoenzymes in biology. *Biochem Soc Trans* **45**, 537–544.
- 10 Ribeiro AJM, Das S, Dawson N, Zaru R, Orchard S, Thornton JM, Orengo C, Zeqiraj E, Murphy JM & Evers PA (2019) Emerging concepts in pseudoenzyme classification, evolution, and signaling. *Sci Signal* **12**, eaat9797.
- 11 Jeffery CJ (2019) The demise of catalysis, but new functions arise: pseudoenzymes as the phoenixes of the protein world. *Biochem Soc Trans* **47**, 371–379.
- 12 Jeffery CJ (2014) An introduction to protein moonlighting. *Biochem Soc Trans* **42**, 1679–1683.
- 13 Force A, Lynch M, Pickett FB, Amores A, Yan YL & Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545.
- 14 Des Marais DL & Rausher MD (2008) Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature* **454**, 762–765.
- 15 Ohno S (1970) Evolution by Gene Duplication. Springer-Verlag, Berlin.



- 16 Wichelecki DJ, Balthazor BM, Chau AC, Vetting MW, Fedorov AA, Fedorov EV, Lukk T, Patskovsky YV, Stead MB, Hillerich BS *et al.* (2014) Discovery of function in the enolase superfamily: D-mannonate and D-gluconate dehydratases in the D-mannonate dehydratase subgroup. *Am Chem Soc* **53**, 2722–2731.
- 17 Pils B & Schultz J (2004) Inactive enzyme-homologues find new function in regulatory processes. *J Mol Biol* **340**, 399–404.
- 18 Pils B & Schultz J (2004) Evolution of the multifunctional protein tyrosine phosphatase family. *Mol Biol Evol* **21**, 625–631.
- 19 Todd AE, Orengo CA & Thornton JM (2002) Sequence and structural differences between enzyme and non-enzyme homologs. *Structure* **10**, 1435–1451.
- 20 Murphy JM, Mace PD & Evers PA (2017) Live and let die: insights into pseudoenzyme mechanisms from structure. *Curr Opin Struct Biol* **47**, 95–104.
- 21 The UniProt Consortium (2018) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47**, D506–D515.
- 22 Ribeiro AJM, Holliday GL, Furnham N, Tyzack JD, Ferris K & Thornton JM (2018) Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites". *Nucleic Acids Res* **46**, D618–D623.
- 23 McDonald AG & Tipton KF (2014) Fifty-five years of enzyme classification: advances and difficulties. *FEBS J* **281**, 583–592.
- 24 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25–29.
- 25 The Gene Ontology Consortium (2018) The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res* **47**, D330–D338.
- 26 Giglio M, Tauber R, Nadendla S, Munro J, Olley D, Ball S, Mitraka E, Schriml LM, Gaudet P, Hobbs ET *et al.* (2018) ECO, the evidence & conclusion ontology: community standard for evidence information. *Nucleic Acids Res* **47**, D1186–D1194.
- 27 El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A *et al.* (2018) The Pfam protein families database in 2019. *Nucleic Acids Res* **47**, D427–D432.
- 28 Potter SC, Luciani A, Eddy SR, Park Y, Lopez R & Finn RD (2018) HMMER web server: 2018 update. *Nucleic Acids Res* **46**, W200–W204.
- 29 Katoh K & Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772–780.
- 30 Capella-Gutiérrez S, Silla-Martínez JM & Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973.
- 31 Price MN, Dehal PS & Arkin AP (2010) FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490.
- 32 Federhen S (2012) The NCBI Taxonomy database. *Nucleic Acids Res* **40**, D136–D143.
- 33 Stolzer M, Lai H, Xu M, Sathaye D, Vernot B & Durand D (2012) Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics* **28**, i409–i415.
- 34 Talevich E, Invergo BM, Cock PJ & Chapman BA (2012) Bio. Phylo: a unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinformatics* **13**, 1–9.
- 35 Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423.
- 36 Huerta-Cepas J, Serra F & Bork P (2016) ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol* **33**, 1635–1638.
- 37 Kreft L, Botzki A, Coppens F, Vandepoele K & Van Bel M (2017) PhyD3: a phylogenetic tree viewer with extended phyloXML support for functional genomics data visualization. *Bioinformatics* **33**, 2946–2947.
- 38 Voadlo DJ, Davies GJ, Laine R & Withers SG (2001) Catalysis by hen egg-white lysozyme proceeds via a covalent intermediate. *Nature* **412**, 835–838.
- 39 Kawaguchi Y, Yoneda K, Torikata T & Araki T (2015) Asp48 function in the hydrogen-bonding network involving Asp52 of hen egg-white lysozyme. *Biosci Biotechnol Biochem* **79**, 196–204.
- 40 Ose T, Kuroki K, Matsushima M, Maenaka K & Kumagai I (2009) Importance of the hydrogen bonding network including Asp52 for catalysis, as revealed by Asn59 mutant hen egg-white lysozymes. *J Biochem* **146**, 651–657.
- 41 Furnham N, Holliday GL, de Beer TA, Jacobsen JO, Pearson WR & Thornton JM (2013) The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res* **42**, D485–D489.
- 42 Sillitoe I, Dawson N, Lewis TE, Das S, Lees JG, Ashford P, Tolulope A, Scholes HM, Senatorov I, Bujan A *et al.* (2018) CATH: expanding the horizons of structure-based functional annotations for genome sequences. *Nucleic Acids Res* **47**, D280–D284.