W60–W66 *Nucleic Acids Research, 2021, Vol. 49, Web Server issue* *Published online 8 May 2021*
https://doi.org/10.1093/nar/gkab328

BENZ WS: the Bologna ENZyme Web Server for four-level EC number annotation

Davide Baldazzi[1], Castrense Savojardo[1], Pier Luigi Martelli[1,*] and Rita Casadio[1,2]

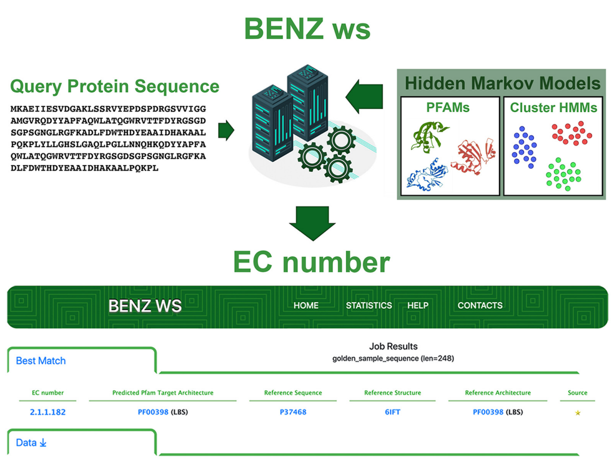
[1]Biocomputing Group, Department of Pharmacy and Biotechnologies, University of Bologna, Italy and [2]Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies (IBIOM), Italian National Research Council (CNR), Bari, Italy

Received March 03, 2021; Revised April 01, 2021; Editorial Decision April 15, 2021; Accepted April 20, 2021

ABSTRACT

The Bologna ENZyme Web Server (BENZ WS) annotates four-level Enzyme Commission numbers (EC numbers) as defined by the International Union of Biochemistry and Molecular Biology (IUBMB). BENZ WS filters a target sequence with a combined system of Hidden Markov Models, modelling protein sequences annotated with the same molecular function, and Pfams, carrying along conserved protein domains. BENZ returns, when successful, for any enzyme target sequence an associated four-level EC number. Our system can annotate both monofunctional and polyfunctional enzymes, and it can be a valuable resource for sequence functional annotation.

GRAPHICAL ABSTRACT



INTRODUCTION

In the post genomic era, annotating protein sequences with functional and structural features is a basic operation for bridging the gap among the hundred millions chains from different organisms, made available by deep sequencing and proteomic projects, and the much smaller number of proteins known with atomic details and with an experimentally characterised biochemical function (1, 2). The problem of functional annotation is therefore one of outmost relevance for the correct assignment of newly generated sequences to their structural and functional protein family or clan, from where they can gain some structural and functional characteristics. Indeed, the experiment Critical Assessment of Functional Annotation (CAFA) (3), since 2010, provides a large-scale assessment of computational methods developed to predict protein function as described with Gene Ontology (GO) terms, according to the three main categories, Molecular Function, Biological Process and Cellular Component (4). Yet, CAFA has no specific section on the Enzyme Commission number (EC number) prediction.

For protein enzymes, the EC number is a traditional code of the catalysed biochemical reactions, describing the relationship among the protein activity, substrates, and products. Presently, ENZYME (5) is the repository of information relative to the nomenclature of enzymes, based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (https://web.expasy.org/docs/swiss-prot_guideline.html). Rhea (6), in turn, is the expert-curated knowledgebase of chemical and transport reactions of biological interest, based on the chemical dictionary ChEBI, which describes reaction participants and their transformations (https://www.rhea-db.org/). In Rhea, reactions are extensively curated with links to supporting literature and are mapped to other resources, including the UniProt file of each protein enzyme. Presently, the EC code includes

*To whom correspondence should be addressed: Tel: +39 0512094005; Fax: +39 0512094005; Email: pierluigi.martelli@unibo.it
Present addresses:
Davide Baldazzi, Biocomputing Group-FABIT, University of Bologna, Bologna 40126, Italy. davide.baldazzi8@unibo.it
Castrense Savojardo, Biocomputing Group-FABIT, University of Bologna, Bologna 40126, Italy. castrense.savojardo2@unibo.it
Pier Luigi Martelli, Biocomputing Group-FABIT, University of Bologna, Bologna 40126, Italy. pierluigi.martelli@unibo.it
Rita Casadio, Biocomputing Group-FABIT, University of Bologna, Bologna 40126, Italy. rita.casadio@unibo.it

seven major classes: (i) oxidoreductases, (ii) transferases, (iii) hydrolases, (iv) lyases, (v) isomerases, (vi) ligases, (vii) translocases. The EC code may range from one to four figures, when the protein catalytic activity is characterised with atomic resolution. In this case, when possible, the architecture of the catalytic site is derived from the protein structure and archived in specific databases, like M-CSA (https://www.ebi.ac.uk/thornton-srv/m-csa) (7), which also includes ligands.

In the UniProt reference database for protein sequences, the annotation of a protein as an enzyme is carried out whenever the automated workflow highlights specific features according to given rules (https://www.uniprot.org/help/biocuration). The system implements motifs derived from HAMAP (High-quality Automated and Manual Annotation of Proteins, https://hamap.expasy.org/), (8) and/or PROSITE, (a database of protein domains, families and functional sites, https://prosite.expasy.org/), (9). Feature discovering includes also the presence of motifs described in InterPro (10), which provides functional analysis of proteins by classifying them into families and by predicting domains and important sites (https://www.ebi.ac.uk/interpro/), and in Pfam (11), which models protein families with Hidden Markov Models (HMMs) after multiple sequence alignment (https://pfam.xfam.org/). Via transfer of knowledge and association rules, the enzyme gains an EC number. Eventually, manual curation allows the enzyme sequence to move from the TrEMBL to the SwissProt section of UniProt (https://www.uniprot.org/). EC number annotation in UniProt can include from one to four numbers, routinely depending on the annotation level of the target protein.

Other databases, by integrating different sources of information, comprising UniProt and PDB, offer a complete annotation for enzymes, such as BRENDA, (12), (https://www.brenda-enzymes.org/index.php) and CATH (https://http://www.cathdb.info/) (13). BRENDA, established in 1987, has evolved into a main collection of curated functional enzyme and metabolism data, supported by links to literature and continuously updating (12). CATH, in turn, is a free, publicly available online resource that provides information on the evolutionary relationships of protein domains. Created in the mid-1990s, it is also continuously updated. In its section FunFams, it allows the search of a target sequence and returns a functional annotation with EC number, after protein domain annotation modelled by a system of HMM hierarchical architectures. CATH is also part of InterPro and contributes therefore to the main annotation system of UniProt (https://www.ebi.ac.uk/interpro/).

As an alternative to transfer of knowledge, *ab-initio* computational approaches can give direct prediction/annotation of an EC number for a given input sequence or structure. This approach requires exploring the complex rules of associations among enzyme sequential and structural features and the EC codes. Methods, mainly based on different types of statistical and machine learning methods, adopt different input features, and predict EC numbers ranging from one to four levels, although with an efficiency decreasing at increasing number of levels (for an extensive review, see (14)). More recently, ECPred (15) implements an ensemble of machine learning methods based on EC nomenclature and outperforms DEEPre, based in turn on an end-to-end feature selection and a classification model training approach (16). Both methods declare a decrease in efficiency when predicting four-level EC numbers.

A major problem in annotating EC codes remains their specificity (four-level EC codes) and the EC assignment to polyfunctional enzymes. Here, to address this problem, we develop BENZ, a system including two main sets of HMMs. One set is meant to detect sequence conservation of the target towards functional families, and the other conservation of structural architectures and family domains as described by Pfam models. The information derived from the interplay of the two different types of HMMs allows, in our case, a direct prediction of a four-level EC code for monofunctional enzymes. The system can also associate four-level EC codes to polyfunctional enzymes.

MATERIALS AND METHODS

Databases

BENZ is presently updated with UniProt/SwissProt release 2021_01. A previous version of BENZ, based on UniProt/SwissProt release 2019_11 was used in order to generate a system for CAFA-like validations. Links to Pfam (11) and KEGG (17) databases are derived from the UniProt releases. Fragments and sequences shorter than 50 residues are not considered. Annotations of active, metal, ligand-binding sites (when available) are also derived from UniProt and mapped into the Pfam architecture of the enzyme proteins.

Graph building, clustering and cluster HMM generation

The procedure stands out from a previously implemented workflow, which we adopted to generate and update our BAR 3.0 (Bologna Annotation Resource, https://bar.biocomp.unibo.it/bar3/), a protein functional and structural annotation resource (18). Briefly, all the UniProt sequences of a specific release (in this case, UniProtKB 2019_01) are compared with BLAST (https://www.ncbi.nlm.nih.gov/), and then clustered by constraining sequence identity (SI) and alignment coverage (COV, the ratio between the number of overlapping positions and the alignment length). A graph is built by connecting sequence pairs that fulfil both identity and coverage constraints. Here, we adopt (SI) $\geq$50% on an alignment coverage (COV) $\geq$90%. Clusters are obtained by isolating the connected components of the graph. For updating, we use UniRef90 clusters (https://www.uniprot.org/help/uniref) which are mapped to BAR clusters, following the procedure outlined before (18). This allows the inclusion of the remaining TrEMBL sequences, and the AlignBucket algorithm (20) speeds up the alignment procedure, exploiting the constraint on COV. Each sequence in the cluster retains the annotation present in the UniProt file (PDB with the highest coverage and resolution when available, Pfam/s, KEGG links and four-level EC codes, when present). Our system allows updating (18), by adding new sequences and by reshaping clusters accordingly, with the inclusion of new annotations from UniProt.

From this background architecture, we retain only clusters containing sequences associated to four-level EC codes, particularly clusters containing SwissProt manually curated sequences, and TrEMBL sequences with an associated PDB file. For each cluster, we then trained a cluster HMM, with HMMER 3.3.2 (http://hmmer.org/, (20)), on the cluster specific multiple sequence alignment, as computed with Clustal Omega (21). The present version of BENZ WS, for technical reasons includes Cluster HMMs with average lengths ranging from 50 to 5000 residues, and this sets the limit of the query sequence to about 5000 residues.

Reference sequence selection and cluster HMM coloring scheme

For each cluster HMM, we select the best annotated sequence/s to be *reference sequence* for the cluster HMM-EC number/s association with the following constraints: for SwissProt sequences, chains with the highest annotation score; for TrEMBL sequences, only those with a four-level EC number and a PDB association. Each reference sequence is then associated to its specific Pfam architecture, and eventually relevant sites (including active, ligand and metal binding sites) are mapped into the corresponding Pfam/s. Cluster HMM are then grouped into two categories. GOLD cluster HMMs are univocally associated to one reference sequence, and BLUE cluster HMMs are associated to more than one reference sequence.

BENZ implementation

BENZ includes cluster HMMs and Pfam models (Pfam version 33.1). When a target sequence enters the server, it is filtered by the two different sets of models. Within the cluster HMMs, when retained (threshold for inclusion is *E*-value $\leq 10^{-5}$), the sequence finds a reference template; within the Pfam models, when retained (threshold for inclusion is E-value$\leq 10^{-4}$), it gains an architecture. The inclusion E-values were chosen after a self-consistency test (the prediction of the whole set of reference sequences).

This architecture is then compared to that of the reference and the target is endowed with the four-level reference EC number only when its architecture is at least equal to that of the template. If not, the four-level EC number is attributed on the basis of a common Pfam, containing relevant sites (active, metal binding, ligand binding sites). The general scheme of BENZ annotation system is depicted in Figure 1. When the retaining cluster HMM is plurivocally associated to more than one reference sequence, a dendrogram is generated after multiple sequence alignment with Clustal Omega (21), including the query sequence, which is associated with the EC code/s of the most similar among the references.

Web server

The BENZ web server interface is optimized to work with common web browsers, including Chrome 88.0, Firefox 83.0, Edge 88.0 and Safari 14.0. Upon submission, jobs are processed asynchronously adopting an internal queuing service based on Sun Grid Engine. Submitted sequences are

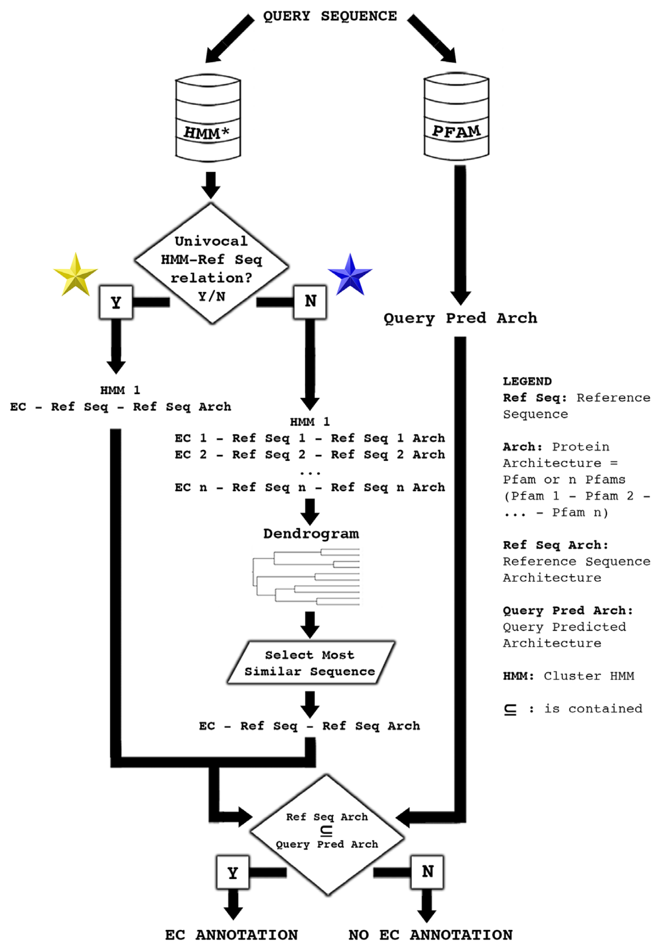


Figure 1. Workflow of BENZ WS. For a query sequence, in FASTA format, the annotation procedure starts with HMM filtering. If the retaining HMM is plurivocally associated to different references sequences (blue star), a dendrogram is generated to find among the reference sequences the most similar one to the target. Otherwise (yellow star), the target is associated to the only reference. The EC number-query sequence association is then made after evaluating if the reference protein architecture (Ref Seq Arch) is contained (⊆) in that of the predicted target Pfam architecture (Query Pred Arch), focusing on Pfams carrying relevant sites. Pfams in our system are annotated when possible, with the positions of the active site, ligand binding site and metal binding site (relevant sites). A sequence feature viewer allows the user to verify whether the query sequence conserves the residues relevant to the protein catalysis for validating the transfer of annotation from the reference sequence. Links to the reference sequence UniProt/SwissProt file, structure PDB file and Pfam entries, together with KEGG identifiers and pathways are also present in the output (see HELP, https://benzdb.biocomp.unibo.it/help).

aligned against the cluster HMMs and Pfam libraries with HMMer 3.3.2 (20). User is provided with a link to a static web page that will display results upon job completion. The page is updated every 30 s. Results are routinely returned within 1 minute since the submission. Longer times may be needed for sequences longer than 3000 residues.

When criteria described in Figure 1 are fulfilled, the result page returns the EC annotation as derived from the best matched reference sequence. The 'Best match' section also reports the PDB structure of the reference (when available), the Pfam architectures of both query and reference sequences and the type of HMM-reference association, either univocal (GOLD star) or plurivocal (BLUE star). More details are provided in the 'Data' section, including the list of cluster HMMs scoring with *E*-value $\leq 10^{-5}$, the associated reference sequences and the links to IntEnz (https://www.ebi.ac.uk/intenz/), UniProt, PDB, Pfam and KEGG. Tabular data are represented with DataTables (https://datatables.net/), allowing to sort rows with respect to any column key and to search for text occurrences in the table. Links are resolved with the Identifiers.org service (22) to improve interoperability.

For plurivocal clusters, the dendrogram, in Newick format, representing the distances among the query and the reference sequences is computed with Clustal Omega (21) and visualized with the Bio.Phylo module of Biopython (23). The Pfam domains mapped on the query sequence are listed in the 'Predicted Target Architecture' table and graphically represented with tracks displayed by means of the Pviz.js library (24). Graphic view also enables to investigate the conservation of active, metal-binding, and ligand-binding sites between the query and the reference sequences. The web server is freely accessible without registration at https://benzdb.biocomp.unibo.it.

RESULTS

BENZ statistics

In the present version, our annotation system comprises 16 593 reference sequences (93.6% from SwissProt), from 891 organisms, included in 12 612 cluster HMMs (Table 1). Our system can annotate 5136 four-level EC numbers by means of a target-reference sequence association (Figure 1). This can be found by filtering the target with cluster HMMs and by associating the predicted target architecture to that of the reference sequence (Figure 1). When more than one reference is present in the retaining cluster HMM, a dendrogram, including the target and the cluster HMM references, allows finding the closest reference to the target. The final comparison among the predicted target architecture and the reference selected one, allows or not the association of the target with the EC number of the reference. In BENZ, 16% of the clusters HMMs (BLUE) are endowed with more than one reference sequence, including 6798 reference sequences (36% of the references, Table 1).

The reference sequence architectures comprise 4158 Pfam models, 1758 of which map relevant sites (active, ligand and metal binding) for testing the target vs reference conservation of the functional activity. 7601 reference sequences are linked to 9382 KEGG pathways.

BENZ comprises also 2023 polyfunctional reference sequences (96% from SwissProt), for a total of 1589 four-level EC numbers, included in 1485 cluster HMMs (907 GOLD and 578 BLUE). The distribution of the polyfunctional reference sequences (Supplementary Table S1S and Supplementary Figure S1S) indicates that the number of EC codes per sequence ranges from 2 to 9, following the UniProt annotation. Their associated architecture includes from one to 26 Pfam models, for a total of 1156 Pfam entries. Polyfunctional reference enzymes have relevant sites mapped into 725 Pfam entries and 1082 polyfunctional reference sequences link 2627 KEGG pathways.

BENZ at work

BENZ is tested against different protein sets (Table 2). Firstly, we run two different sets of proteins not included in our reference sequences: a positive (sequences annotated in SwissProt with a four-level EC code) and negative (sequences annotated in SwissProt without a four-level EC code). Results indicate that the system has a good efficiency in assigning four-level EC codes (92.4%) and in rejecting non-enzyme proteins (95.1%). A similar good efficiency is detected when testing two other sets, one comprising polyfunctional enzymes from SwissProt and the other including human sequences endowed with EC numbers downloaded from TrEMBL.

CAFA-like validation

The performance assessment of our method was carried out running an in-house CAFA-like benchmark. To this aim, we simulated a time-challenge experiment by computing EC annotation acquired in the time elapsed between two distant releases of SwissProt. As reference sets, we used SwissProt releases 2019_11 (*t0*: 11 December, 2019) and 2020_03 (*t1*: 17 June 2020). A BENZ test version was implemented using only sequences and annotations of the former release. Positive and negative benchmark datasets were compiled by comparing the functional annotations available in the two releases. The positive dataset consists of proteins non-annotated for EC at *t0* but endowed with a four-level EC annotation at *t1*. Fragments were excluded. The full positive dataset therefore consists of 607 proteins not included in the ground-truth dataset of the BENZ-WS test version and endowed with a four-level EC number out of the seven main EC classes. For sake of comparison with methods not handling the EC 7 class (translocases), we considered a reduced dataset comprising 366 enzyme sequences labelled with EC codes from classes 1 to 6.

The negative dataset contains 1034 non-fragment proteins that, from *t0* to *t1*, acquired a Gene Ontology (GO) annotation for Molecular Function (MF) different from GO:0003824 (catalytic activity) and its descendants, and that are not endowed with an EC number at any level.

We then assessed the performance of the BENZ testing version (built only on sequences and annotations available at *t0*) in discriminating enzymes from other proteins and in assigning the EC annotation. We computed different scoring measures, including the True Positive Rate (TPR), evaluating the fraction of correct predictions at different EC levels, the False Negative Rate (FNR) scoring the number of enzymes in the positive dataset predicted as non-enzymes and the False Positive Rate (FPR) scoring the number of negative proteins predicted with an EC number (Table 3).

On the full dataset (Table 3, first row), BENZ reaches FNR and FPR values of 12.2% and 3%, respectively, indicating a good ability in discriminating enzymes from other proteins. The correct EC number assignment (TPR) is equal to 85% on four-level annotations, and slightly higher when less detailed levels of annotation are considered. When the seventh Enzyme class is filtered out in the reduced data set (about 40% of the proteins), BENZ WS is still scoring with good values of FNR and FPR (second row in Table 3), highlighting the robustness of the method.

Table 1. BENZ statistics

	EC 1	EC 2	EC 3	EC 4	EC 5	EC 6	EC 7	Total
EC numbers[a]	1437	1550	1034	595	254	189	77	5136
Cluster HMM	1758	5116	3755	1006	637	636	288	12 612
Cluster HMM GOLD	1326	4315	3190	800	497	496	218	10 547
Cluster HMM BLUE	432	801	565	206	140	140	70	2065
Ref Seq[b]	2752 (390)	6455 (990)	4582 (729)	1348 (324)	842 (145)	883 (105)	405 (14)	16 593 (2023)
Ref Str[b]	1230 (152)	2252 (253)	2100 (213)	625 (110)	333 (41)	287 (22)	149 (5)	6798 (618)
Pfam[c]	682 (429)	1923 (769)	1672 (711)	463 (321)	294 (190)	276 (133)	143 (50)	4158 (1758)
KEGG ID[d]	2390	5908	3770	1185	799	894	343	14 745
KEGG Pathway[d]	2758	4812	2628	1972	952	1266	317	9382
Organisms[e]	15 158 200 13 Arc Bac Euk Vir	20 187 232 1 193 Arc Bac Euk Unk Vir	15 165 208 1 135 Arc Bac Euk Unk Vir	13 125 141 4 Arc Bac Euk Vir	16 109 83 6 Arc Bac Euk Vir	18 119 53 12 Arc Bac Euk Vir	7 57 47 Arc Bac Euk	24 261 391 2 213 Arc Bac Euk Unk Vir

[a]Four-level EC numbers are distributed according to the 7 EC classes: EC1-Oxidoreductases, EC2-Transferases, EC3-Hydrolases, EC4-Lyases,EC5-Isomerases, EC6-Ligases, EC7-Translocases.
[b]Ref Seq and Ref Str: number of reference sequences, and reference sequence with structure, respectively; number of polyfunctional enzymes are within brackets.
[c]Pfam: models from Pfam (https://pfam.xfam.org); within brackets Pfams, where relevant sites (active, metal, ligand binding site) are annotated.
[d]KEGG ID: from UniProt annotation; KEGG pathway: from https://www.genome.jp/kegg/.
[e]number of organisms detailed for each kingdom. Arc: Archaea; Bac: Bacteria; Euk: Eukaryota; Oth: Others; Vir: Viruses. Unk: unknown. Annotation source: UniProt. Grand Total: 891.

Table 2. BENZ at work

Dataset	Sequences (#)	Acc[e] (%)	FNR[f] (%)	FPR[g] (%)
Positive[a]	197 880	92.4	3.9	-
Negative[b]	12 315	95.1	-	4.9
Polyfunctional[c]	10 764	93.7	5.0	-
TrEMBL-human[d]	10 024	93.4	5.6	-

[a]Positive: the positive set contains complete SwissProt sequences without any PDB counterpart and annotated with only four-level EC number.
[b]Negative: the negative set comprises complete SwissProt sequence with a PDB counterpart, without EC codes.
[c]Polyfunctional: the set includes complete SwissProt sequence that are annotated with two or more four-level EC numbers.
[d]TrEMBL-human: the set contains complete TrEMBL sequences from *Homo sapiens* annotated with a four-level EC number.
[e] Acc (Accuracy) measures the number of proteins correctly assigned. For sets containing positive examples, it corresponds to the True Positive Rate as evaluated at the level of four: EC annotation. For the negative set, it corresponds to the True Negative Rate.
[f]FNR (False Negative Rate) measures the percentage of enzymes predicted as non-enzymes.
[g]FPR (False Positive Rate) measures the percentage of non-enzymes predicted as enzymes.

We then compared BENZ with three state-of-the-art tools: ECPred (15), DEEPre (16) and EFICAz2.5 (25). Only the reduced positive dataset was adopted since the selected methods do not consider the seventh EC class. Results indicate that BENZ outperforms the other tools in this benchmark (Table 3). TPR values of BENZ WS range from two-fold up to four-fold those obtained by the other predictors, increasing at increasing levels of EC. Concomitantly, BENZ WS FNR values overpass other predictors values by at least two or three times those of the other predictors (Table 3, FNR column).

In the reduced set, BENZ achieves a better discrimination than the other methods (FNR) and a better EC assignment sensitivity, with a TPR value ranging from 79.2% to 75% at increasing level of predicted EC (Table 3, TPR columns), and it significantly overpasses the second best-scoring method (DEEPre, 16). As to the correct recognition of non-enzymes (column FPR, Table 3), DEEPre and EFICAz2.5.1 show a better performance, which in turn is counter-balanced by a low ability to recognise enzymes (TPR values).

DISCUSSION

A major problem in addressing EC code annotation is due to the different levels of specificity that the code carries. Only the complete four-level annotation fully characterises the protein biochemical activity. However, due to evolution, different active site architectures can catalyse the same biochemical activity and/or the same active site can bind different substrates (26). These difficulties may hamper the EC direct association to the protein sequence and rather suggest a direct prediction of GO terms, like in the CAFA experiments (3).

Here, we tackled the problem of the association of protein sequence with four-level EC code/s taking advantage of two different types of HMMs. One, the cluster HMM derives from a hierarchical clustering procedure that we adopted before for generating a system (BAR 3.0) suited to a general-purpose protein annotation and based on a rigorous and statistically validated transfer of annotation. Cluster HMMs model sequences, which have been clustered after constraining their identity (≥50%) over 90% of the alignment length. By this, cluster HMMs retain sequences that pairwise share a high level of similarity over a large portion of the alignment length, although belonging to different organisms. Furthermore, they may conserve relevant sites in specific Pfam domains. Among the cluster-sequences, we select one reference sequence (the one with the highest score of annotation) and define its architecture by mapping Pfam domains to the chain. When present, all the relevant sites (active, ligand and metal binding) are also mapped to the corresponding Pfam domain/s. Finally, we associate each representative, its architecture and EC code/s to a more general representation, casted into the cluster HMM. Indeed, structural matching for gaining the EC code of the representative reference is checked by comparing the target predicted architecture and the reference one.

Testing BENZ on selected sets of proteins (Table 2) indicates that the system correctly rejects (97%) non enzymes

Table 3. BENZ benchmarking

Method	Data set[a]	TPR[f] (%) 1 level	TPR[f] (%) 2 level	TPR[f] (%) 3 level	TPR[f] (%) 4 level	FNR[g] (%)	FPR[h] (%)
BENZ WS[b]	Full	87.5	87.5	87.5	85.0	12.2	3.0
BENZ WS[b]	Reduced	79.2	79.2	79.2	75.1	20.2	3.0
ECPred[c]	Reduced	43.7	34.7	23.8	13.1	45.6	12.2
DEEPre[d]	Reduced	38.8	35.2	27.9	20.8	51.1	2.4
EFICAz2.5.1[e]	Reduced	33.6	33.1	31.1	16.7	63.7	1.6

[a]Benchmark datasets are extracted by comparing SwissProt releases 2020_3 and 2019_11. The full dataset includes 607 proteins that have gained EC annotation (7 EC classes); the reduced dataset includes a subset of 366 enzyme sequences without EC codes of the seventh class for comparing with the other predictors. Both datasets comprise 1013 non-enzyme sequences as negative examples.
[b]A BENZ WS version including only sequences and annotations available in the SwissProt release 2019_11 has been used for this test.
[c]ECPred (15) has been downloaded from https://github.com/cansyl/ECPred and run in-house; it does not provide multiclass predictions and the best match between the output and the list of EC numbers has been considered for multiclass enzymes. It does not include enzymes of for EC class 7.
[d]DEEPre (16) predictions have been run on the webserver http://www.cbrc.kaust.edu.sa/DEEPre/ in modality 'I'm not sure the sequence is an enzyme'; it does not provide multiclass predictions and the best match between the output and the list of EC numbers has been considered for multiclass enzymes. It does not include enzymes of the EC class 7.
[e]EFICAz2.5.1 (25) has been downloaded from https://sites.gatech.edu/cssb/eficaz2-5/ and run in-house; it does not include enzymes of EC class 7.
[f]TPR (True Positive Rate) measures the number of enzymes assigned to the correct EC class. TPRs have been evaluated at the level of four-level EC annotation.
[g]FNR (False Negative Rate) measures the percentage of enzymes predicted as non-enzymes.
[h]FPR (False Positive Rate) measures the percentage of non-enzymes predicted as enzymes.

and that it is efficient in retaining never seen before enzyme sequences (Table 3). BENZ will eventually assign only EC codes present in the system as specific four-level EC-Cluster HMM-reference sequence association. This will be taken care of with new BENZ releases, following new UniProt releases.

When BENZ is benchmarked with other EC predictors, based on first structural principles or machine and deep learning methods, it is superior (Table 3). Predictors, which we found available, are based on different methods and not directly comparable. However, their poor performance on the specific task of EC code prediction, including polyfunctional enzymes, suggests that fine-tuning of the protein functional family representation is necessary and that machine learning, including end-to-end models, poorly captures it.

We introduce BENZ as a reliable method for transfer of knowledge after generalisation over subsets of proteins belonging to specific functional and structural families.

DATA AVAILABILITY

BENZ WS is freely available as a web server at the following URL: https://benzdb.biocomp.unibo.it/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

D. Baldazzi is the recipient of a PhD fellowship for the Data Science and Computation PhD program of the University of Bologna, Italy, supported by Centro di Riferimento Oncologico (CRO), Aviano, Italy.

FUNDING

Italian Ministry of Education, University and Research [PRIN2017 grant, project no. 2017483NH8_002 to C.S.]; European Commission H2020 programme [CIRCLES project, grant no. 818290 to P.L.M.]. Funding for open access charge: Italian Ministry of Education, University and Research.

Conflict of interest statement. Not declared.

REFERENCES

1. The UniProt Consortium. (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
2. Burley,S.K., Bhikadiya,C., Bi,C., Bittrich,S., Chen,L., Crichlow,G.V., Christie,C.H., Kenneth Dalenberg,K., Di Costanzo,L. *et al.* (2021) RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.*, **49**, D437–D451.
3. Zhou,N., Jiang,Y., Bergquist,T.R., Lee,A.J., Kacsoh,B.Z., Crocker,A.W., Lewis,K.A., Georghiou,G., Nguye,H.N. and Friedberg,I. (2019) The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome biology*, **20**, 244.
4. Gene Ontology Consortium. (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.*, **49**, D325–D334.
5. Pundir,S., Onwubiko,J., Zaru,R., Rosanoff,S., Antunes,R., Bingley,M., Watkins,X., O'Donovan,C. and Martin,M.J. (2017) An update on the Enzyme Portal: an integrative approach for exploring enzyme knowledge. *Protein Eng. Des. Sel.*, **30**, 245–251.
6. Lombardot,T., Morgat,A., Axelsen,K.B., Aimo,L., Hyka-Nouspikel,N., Niknejad,A., Ignatchenko,A., Xenarios,I., Coudert,E. and Bridge,A. (2019) Updates in Rhea: SPARQLing biochemical reaction data. *Nucleic Acids Res.*, **47**, D596–D600.
7. Ribeiro,A.J.M., Holliday,G.L., Furnham,N., Tyzack,J.D., Ferris,K. and Thornton,J.M. (2018) Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res.*, **46**, D618–D623.
8. Pedruzzi,I., Rivoire,C., Auchincloss,A.H., Coudert,E., Keller,G., de Castro,E., Baratin,D., Cuche,B.A., Bougueleret,L. and Bridge,A. (2015) HAMAP in 2015: updates to the protein family classification and annotation system. *Nucleic Acids Res.*, **43**, D1064–D1070.
9. Sigrist,C.J.A., de Castro,E., Cerutti,L., Cuche,B.A., Hulo,N., Bridge,A., Bougueleret,L. and Xenarios,I. (2013) New and continuing developments at PROSITE. *Nucleic Acids Res.*, **41D**, D344–D347.
10. Blum,M., Chang,H-Y., Chuguransky,S., Grego,T., Kandasaamy,S., Mitchell,A., Nuka,G., Paysan-Lafosse,T., Qureshi,M. and Finn,R.D.

(2021) The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.*, **49D1**, D344–D354.
11. Mistry,J., Chuguransky,S., Williams,L., Qureshi,M., Salazar,G.A., Sonnhammer,E.L.L., Silvio Tosatto,S., Paladin,L., Raj,S. and Bateman,A. (2021) Pfam: the protein families database in 2021, *Nucleic Acids Res.*, **49D1**, D412–D419.
12. Chang,A., Jeske,L., Ulbrich,S., Hofmann,J., Koblitz,J., Schomburg,I., Neumann-Schaal,M., Jahn,D. and Schomburg,D. (2021) BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Res.*, **49**, D498–D508.
13. Sillitoe,I., Dawson,N., Lewis,T.E., Das,S., Lees,J.G., Ashford,P., Tolulope,A., Scholes,H.M., Senatorov,I. and Orengo,C.A. (2019) CATH: expanding the horizons of structure-based functional annotations for genome sequences. *Nucleic Acids Res.*, **47D1**, D280–D284.
14. Tan,J.X., Lv,H., Wang,F., Dao,F.Y., Chen,W. and Ding,H. (2019) A survey for predicting enzyme family classes using machine learning methods. *Curr. Drug Targets*, **20**, 540–550.
15. Dalkiran,A., Rifaioglu,A.S., Martin,M.J., Cetin-Atalay,R., Atalay,V. and Doğan,T. (2018) ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. *BMC Bioinformatics*, **19**, 334.
16. From,Li Y., Wang,S., Umarov,R., Xie,B., Fan,M., Li,L. and Gao,X. (2018) DEEPre: sequence-based enzyme EC number prediction by deep learning. *Bioinformatics*, **34**, 760–769.
17. Kanehisa,M., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
18. Profiti,G., Martelli,P.L. and Casadio,R. (2017) The Bologna Annotation Resource (BAR 3.0): improving protein functional annotation, *Nucleic Acids Res.*, **45**, W285–W290.
19. Profiti,G., Fariselli,P. and Casadio,R. (2015) AlignBucket: a tool to speed up 'all-against-all' protein sequence alignments optimizing length constraints. *Bioinformatics*, **31**, 3841–3843.
20. Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.
21. Sievers,F., Wilm,A., Dineen,D., Gibson,T. J., Karplus,K., Li,W. and Higgins,D. G. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.
22. Juty,N., Le Novère,N. and Laibe,C. (2012) Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. *Nucleic Acids Res.*, **40**, D580–D586.
23. Talevich,E., Invergo,B.M., Cock,P.J. and Chapman,B.A. (2012) Bio.Phylo: a unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinformatics*, **13**, 209.
24. Mukhyala,K. and Masselot,A. (2014) Visualization of protein sequence features using JavaScript and SVG with pViz.js. *Bioinformatics*, **30**, 3408–3409.
25. Kumar,N. and Skolnick,J. (2012) EFICAz2.5: application of a high-precision enzyme function predictor to 396 proteomes. *Bioinformatics*, **28**, 2687–2688.
26. Tyzack,D.J., Furnham,N., Sillitoe,I., Orengo,C.M. and Thornton,J.M. (2017) Understanding enzyme function evolution from a computational perspective. *Curr. Opin. Struct. Biol.*, **47**, 131–139.