MINING THE *MYCOBACTERIUM TUBERCULOSIS* FLAVOPROTEOME: A BIOINFORMATIC APPROACH.

Speaker: Raquel Ventura Baños
Director: Milagros Medina Trullenque
Codirector: Marta Martínez Júlvez

JUNE 11, 2025
UNIVERSIDAD DE ZARAGOZA – DPTO. BIOQUÍMICA – BIFI
Master in Biophysics and Quantitative Biotechnology

Index

Abstract **4**

Dictionary **5**

1 Introduction **6**

1.1 Background of the proposed study *6*

1.2 Drug target search in Mycobacterium tuberculosis and its medical relevance *7*

1.3 Data mining and its importance in biology *8*

1.4 The theory behind sequence similarity search methods and multiple sequence alignment *10*

2 Objective **12**

3 Method **13**

3.1 Experimental design *13*

3.2 Workflow *15*

3.3 Parameters *17*

3.4 Programming languages *17*

3.5 Environment and programming tools *17*

3.6 Databases *18*

3.7 Bioinformatic programs *18*

3.8 Equipment *18*

4 Results and discussion **18**

4.1 Functionally uncharacterized flavoproteins *18*

4.2 Method performance *18*

4.2.1 Prediction vs. reference annotations evaluation parameters. 20

4.2.2 Overall predictive precision, recall and ortholog annotation coverage 21

4.2.3 Query predictive precision, recall and ortholog annotation coverage. 21

4.3 Biological interpretation of RBH unknown protein results *30*

4.3.1 Predictions for queries with new information only regarding EC numbers: 30

4.3.2 Predictions for queries with new information regarding pathways: 32

4.3.3 Predictions for queries with new information regarding EC numbers and pathways: 34

4.4 Biological interpretation of UBH unknown protein results *36*

4.4.1 Predictions for queries with new information regarding EC numbers: 36

4.4.2 Predictions for queries with new information regarding pathways: 38

4.4.3 Predictions for queries with new information regarding EC numbers and pathways: 42

4.5 Metabolic context and pharmaceutical and biotechnological applications of newly inferred protein functions. *48*

5 Conclusions **50**

Bibliography **52**

Abstract

Data mining techniques have proven to be really useful in biology. Text mining techniques, such as similarity search, allow us to search for patterns in sequences. The similarity of protein sequences provides insight into function when the correct genetic and statistical considerations are applied. One group of biologically relevant proteins for which such approaches are especially valuable is flavoproteins, a wide group of proteins with diverse and characteristic functionality. They are defined by their ability to bind riboflavin derivatives and participate primarily in electron transfer processes. The catalytic activity of these type of proteins in most species is not fully annotated. Thus, the flavoproteome of *Mycobacterium tuberculosis* is suspected to contain 184 different protein sequences, of which only 51 are fully functionally described.

With the aim of finding new drug targets for the treatment of tuberculosis, an infectious disease caused by this organism, and to explore its new metabolic functions and possible uses, a search was performed for similar flavoproteins to the uncharacterized ones. The method developed in this study is based on the reciprocity of the most significant similarities between species sequences, obtaining orthologs from all taxonomic groups present in a custom database of activity-annotated flavoproteins. All 184 flavoproteins belonging to *M. tuberculosis* were queried in 48 minutes, obtaining more accurate catalytic functions for 33 of them, and identifying 4 possible drug targets and 2 potential biocatalysts. The search output also allowed for a better understanding of method recall and similarity significance by evaluating the coherence of the method's predictions with previous annotations.

Dictionary

BBH: Bidirectional best hits

BLAST: Basic local alignment search tool

DM: Data mining

DS-TB: Drug susceptible tuberculosis

ESKAPE: Acronym comprising the scientific names of six highly virulent and antibiotic resistant bacterial pathogens including: *Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa, and Enterobacter spp*.

FAD: Flavin adenine dinucleotide

FMN: Flavin mononucleotide

HMM: Hidden Markov models

HSPs: High scoring segment pairs

KDD: Knowledge discovery in databases

MDR-TB: Multiple drug resistant tuberculosis

MSA: Multiple sequence alignment

PPI: Protein-Protein Interactions

RBH: Reciprocal best hits

RR-TB: Rifampicin resistant tuberculosis

TB: Tuberculosis

UBH: Unidirectional best hits

1 Introduction

1.1 Background of the proposed study

Original definition of flavoenzymes present them as electron-transferring enzymes that contain a bound flavin prosthetic group; usually flavin mononucleotide (FMN) and/or flavin adenine dinucleotide (FAD) riboflavin derivatives. This is generally true since for most of these enzymes the flavin cofactor works as an electron donor and acceptor for the substrate to become reduced or oxidized (Lienhart et al., 2013; Massey, 2000, Singer and Edmondson, 1978). In general, redox changes in the flavin occur at its isoalloxazine ring, between N atoms 1 and 5 of the flavin isoalloxazine ring (Figure 1) (Massey, 2000; Miura, 2001). In addition, some redox flavoenzymes are more complex systems, since they have additional electron carriers apart from flavins such as Zn^{2+}, $Mo^{+3,+5}$ or Fe-S centers (Singer and Edmondson, 1978).

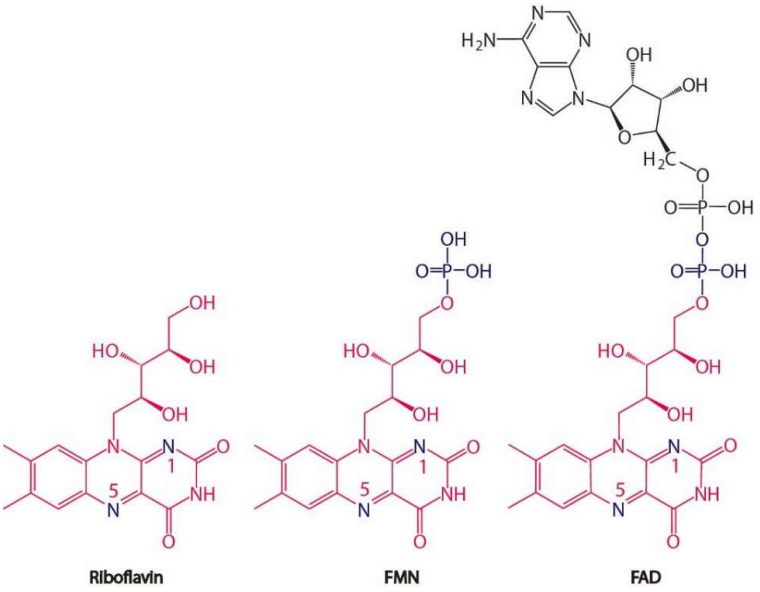

Figure 1. Riboflavin is the precursor of FAD and FMN cofactors, both indispensable cofactors for flavoproteins and flavoenzymes, image taken from Zhang et al., 2020.

Since the flavin function is not only limited to redox processes, a more accurate description of flavoproteins is given when they are simply defined by a protein bound to the flavin cofactor, FMN or FAD (Massey, 1995). Approximately 10% of flavin-dependent enzymes catalyze non-redox reactions. Moreover, the flavin cofactor is also widely used as a signaling and sensing molecule in biological processes such as phototropism and nitrogen fixation. It is estimated that 91% of flavin-dependent enzymes are oxidoreductases, and the remaining enzymes can be classified as transferases (4.3%), lyases (2.9%), isomerases (1.4%) and ligases (0.4%) (Macheroux et al., 2011).

Nonetheless, the exact knowledge of the flavoprotein content in most species is unknown, since the flavoproteome content has only been reported for a small number of species. Fully described flavoproteomes are those in *Homo sapiens, Saccharomyces cerevisiae*, *Arabidopsis thaliana* and *Brucella ovis*. The information revealed in these studies suggests that while some organisms rely heavily on flavoenzyme's activity others remain having less functional diversity regarding flavoproteins. While human and *Saccharomyces cerevisiae* flavoproteomes contain 78 and 48 different proteins respectively, Arabidopsis thaliana has more than 200 (Minjárez-Sáenz *et al.*, 2022; Eggers *et al.*, 2021; Gudipati *et al.*, 2014; Lienhart *et al.*, 2013).

The study addressed in this work arose from a previous analysis aimed to identify all the components of the *Mycobacterium tuberculosis* H37Rv strain flavoproteome (Montesa et al., 2023 TFG), where 184 proteins were identified as potential flavoproteins by searching on NCBI (https://www.ncbi.nlm.nih.gov/), UniProt (https://www.uniprot.org/), BioCyc (https://biocyc.org/) and Mycobrowser (https://mycobrowser.epfl.ch/) Databases. Different classifications for clades and domains were also performed by using Pfam and InterPro databases, as well as virulence factor predictors. After homology search, for each protein it was also evaluated the presence of homologous proteins in mammals (*Homo sapiens* included) and ESKAPE organisms. In addition, BRENDA, KEGG, Mycobrowser and UniProt databases were used to annotate possible functions and metabolic routes (Montesa et al., 2023 TFG).

The study included an attempt to classify in depth enzyme function by the chemical reaction catalyzed using as a standard a number called Enzyme Commission (EC) number that consists of a row of four numbers separated by points. The first EC number indicates if the enzyme can act as a 1: Oxidoreductase, 2: Transferase 3: Hydrolase, 4: Lyase, 5: Isomerase or 6: Ligase. Further numbers indicate with increasing accuracy the kind of substrates the enzyme is dealing with (Hu *et al.*, 2012). Unfortunately, the Montesa´s study lacked the complete identification of the potential function of a relevant number (133) of proteins out of the 184 envisaged as flavoproteins and flavoenzymes. Therefore, possible functions for some of the members of the *M. tuberculosis* flavoproteome were only partially described, and, in many cases, no substrates were identified. Moreover, there were some proteins that could not be identified as flavoproteins because it was not even clear whether they would bind any cofactor, having their sequences as the only guide that may provide information about their activity using the correct tools. Consequently, further research is required (Montesa et al., 2023 TFG).

1.2 Drug target search in *Mycobacterium tuberculosis* and its medical relevance

Tuberculosis (TB) is a worldwide infectious disease caused by *M. tuberculosis* bacteria. It spreads from person to person through airborne transmission. The illness is known to cause approximately 13% of antimicrobial-resistant deaths (Farhat *et al.*, 2024). Treatment for drug-susceptible (DS) TB usually involves taking multiple antibiotic tablets for more than 6 months, while the standard treatment for multiple drug-resistant (MDR) TB is almost 2 years of both oral and injected antibiotics. These toxic regimens frequently cause nausea, liver damage and irreversible hearing loss, and may also require surgery or other invasive procedures. MDR-TB therapy is successful in only around half of cases globally, with a high risk of treatment failure and death (Bark *et al.*, 2024; Nguyen *et al.*, 2020). In addition, multiple drug resistance regimes are not available in every country. The availability of drugs for MDR/RR-TB treatment in Central and Western Europe is shown below (Figure 2) (Otto-Knapp *et al.*, 2024).

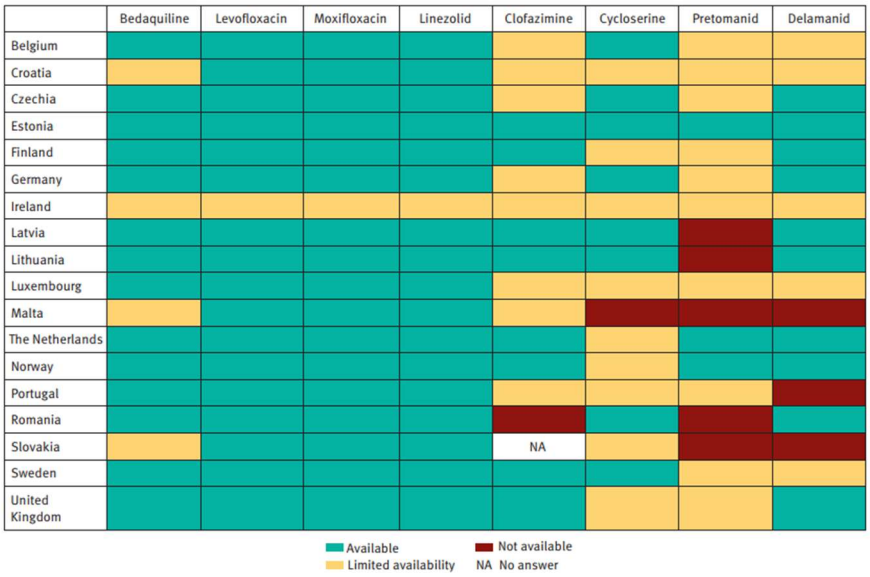

Figure 2. Availability of drugs in different countries. Categorical answers given to the question 'Indicate availability of the following medicines for treatment of MDR/RR-TB in your setting/country', WHO Regional Office for Europe survey, central and western Europe, October 2023 (n = 18 countries), image taken from Otto-Knapp et al., 2024.

The most common way *M. tuberculosis* bacteria acquire resistance to drugs is by altering the target enzyme. Depletion or inactivation of drug-activating enzymes is also common, while upregulation of drug efflux or enzyme inactivation of the drug is less frequent (Farhat *et al.*, 2024).

Due to the need to develop new strategies to address this global health concern, the search for new drugs and targets to treat the disease is an ongoing issue (Global Tuberculosis Report, 2024). Flavoproteins are known to be a wide family of proteins participating in many core functions of the cell and are therefore suitable as potential drug targets (Minjárez-Sáenz *et al.*, 2022; Cremades *et al.*, 2009). Therefore, exploring the flavoproteome of specific bacteria could lead to the identification of new protein targets and the development of new drugs, as well as the use of other known ligands as treatment.

Determining whether the protein at hand might be a good target is not an easy task, and for this concern, previous annotations and predictions are used if available. Protein similarities can also be used to infer the existence of a similar protein in the human core, the involvement in a vital function of the bacteria, enzymatic activity of the protein, structure and possible ligands (Abbasi Mesrabadi et al., 2023; Zhong et al., 2021; Iwata et al., 2013).

1.3 Data mining and its importance in biology

Data mining (DM) can be described as a step of knowledge discovery in databases (KDD); "KDD is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (Frawley et al., 1991). DM is a step in KDD where computational techniques are applied to find patterns in the data (Džeroski, 2008). DM tasks include predictive modeling and descriptive analysis techniques that consist of clustering and summarization. DM is the central step in the KDD process. Other steps involve preparing the data for DM and evaluating the patterns discovered.

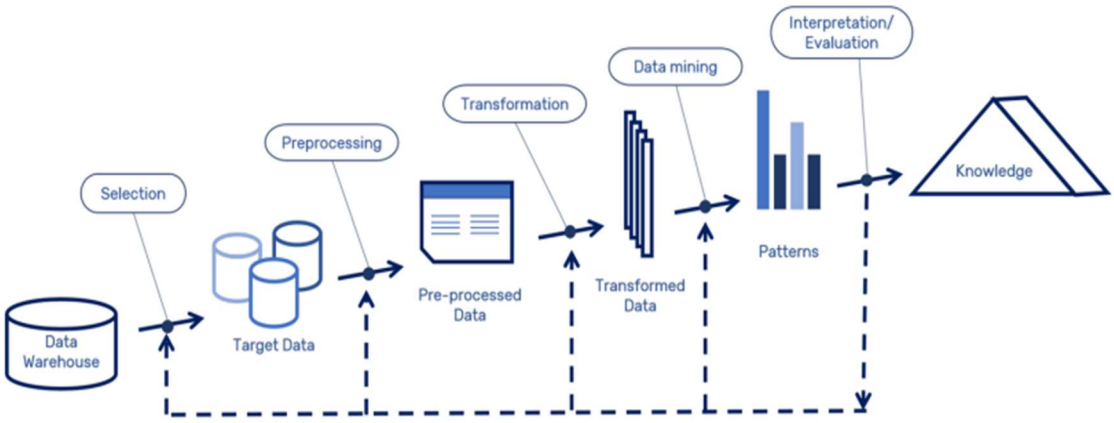

Figure 3. Steps in knowledge discovery in databases (KDD), *image taken from Rotondo and Quilligan, 2020.*

DM has been very useful as a part of data optimization techniques when using DNA, RNA and protein sequences. DM applications on biology vary from outlier removal to DNA and protein structure prediction from sequence (Singh and Singh, 2021), and play a significant role in many biotechnological issues, identifying associations and patterns among data for:

- Alignment and comparison of DNA, RNA and protein sequences
- Gene identification from DNA sequences.
- Understanding of gene expression and microarray data.
- Building of phylogenetic trees for studying evolutionary relationships.
- Protein structure prediction and classification.
- Molecular design and molecular docking.

A very large number of different algorithms have been developed and implemented in the context of data optimization, with clustering being the most common DM technique. This method has been used for data feature removal, sequence clustering, increasing clustering efficiency or removing outliers. Clustering can be performed using a variety of methods, including Partitional and Hierarchical Clustering, K-modes, K-means, Graph Clustering, and K-Nearest Neighbors (Singh and Singh, 2021; Sindhu and Sindhu, 2017).

Another tool widely used in DM for data optimization is “text mining”, which is typically used for identifying incomplete genes and constructing new gene pathways. This specific application of text mining is generally conducted using the Monte Carlo algorithm (Al-Dalky *et al.*, 2016). For data optimization there are also other strategies that are not part of the DM step; namely Genetic Algorithms inspired by human evolution, Swarm Intelligence, Ensemble Learning and Frog Leaping (Thareja and Chhillar, 2020; Hu *et al.*, 2018; Surya Narayana and Vasumathi, 2018; Newman and Cooper, 2010). The last three techniques are artificial intelligence (AI) and Machine Learning algorithms based on the interaction of individual organisms in nature.

DM has been proven also useful in disease prediction, not only for predicting epidemic outbreaks, measuring the efficiency of health programs or identifying a disease at a genome level, but also for exploring disease mechanisms and finding treatments based on protein structure and interactions (Singh and Singh, 2021; Jackson *et al.*, 2018).

Mining protein function has been afforded in a wide variety of ways, with BLAST, based on sequence similarity searches, being the most commonly used tool for this task. (McGinnis and Madden, 2004). Alternatives and complements to this tool include web searching for annotations in different databases (Macheroux et al., 2011), evaluation of conserved subgraphs for protein-protein interactions (PPI) and orthologs (text mining) (Jaeger *et al.*, 2008), protein structure mining (Andreeva *et al.*, 2020), and multiple sequence alignment (MSA) together with phylogenetic analyses (Mansour et al., 2009).

In addition, there are other bioinformatics tools that can predict on structure, structure similarity and protein-protein interactions such as: AlphaFold in its different versions (Jumper *et al.*, 2021), https://www.biorxiv.org/content/10.1101/2024.03.19.585735v1, the DeepRank graph neural network for scoring protein–protein models using protein language model (Xu and Bonvin, 2024), GeoMine (Graef *et al.*, 2022) and other different deep learning approaches for mining protein data (Siu *et al.*, 2010).

1.4 The theory behind sequence similarity search methods and multiple sequence alignment

The homology of two sequences is inferred when sequences share more similarity than the one explained by chance, so the simplest explanation would lead us to think they didn't arise independently but from a common ancestor. Sequences with significant similarity are inferred to be homologous, that is, they may have a common ancestor. Then, if similarity is statistically significant, we can confirm that the sequences are homologous, if not homology is uncertain. Measures of statistical similarity include excess similarity (E-value), identity percentage and bits. The E-value is the significance measure, meaning the quantity of chance-paired homologs found with the similarity retrieved. Very low values are highly significant. Identity percentage depends more on the biological context and refers to the degree of similarity (Pearson, 2013; Kerfeld and Scott, 2011).

Although sequence and structural homology are easy to infer, function and similarity are not always as closely related as expected. Orthologs are similar sequences that share a common ancestor, while paralogs are generated by duplication (figure 4). Given the ortholog conjecture, which states that proteins from orthologs have higher chances to share functionality, orthologs are often used to predict function based on similarity. However, there is ongoing debate about the terms "ortholog" and "paralog", that are often mistakenly used to describe other evolutionary relationships. This has led to the development of related terms to duplication and speciation such as "inparalog", "outparalog", "pseudoortholog", "coortholog"… (Stamboulian et al., 2020; Pearson, 2013; Koonin, 2005).

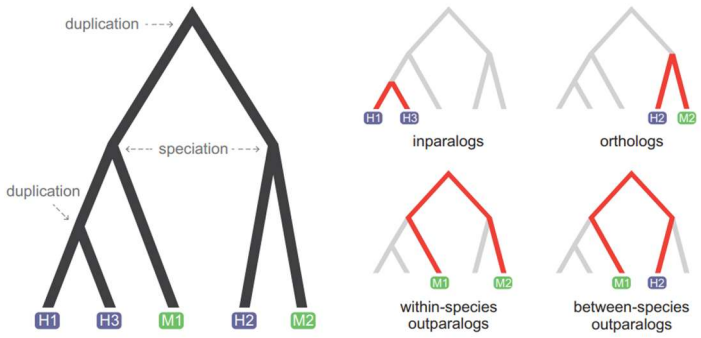

Figure 4. Four types of homology relations. A family of five genes sampled from humans (in blue) and mice (in green) evolves through speciation and duplication events (left-hand tree), images taken from Stamboulian et al., 2020.

The most widely used tool for sequence similarity search, BLAST, allows a matching process between a query and the database sequences (McGinnis and Madden, 2004). BLAST algorithm for protein sequences works with high-scoring segment pairs with short words as seeds. The principal parameters that the algorithm uses are word size, w; similarity threshold, S; and minimum match score, T. First, a list of words that have a similarity to the query sequence higher than the threshold, S, is made. Then, every database sequence is scanned to find the exact matches of the words in the list. For each match, the alignment is expanded across both sides of the sequence to find a total maximum match score, generating a high scoring segment pair HSP (Altschup *et al.*, 1990). Exact string matching for similarity search is performed by algorithms based on Boyer-Moore, Knuth-Morris-Pratt and Aho-Corasick algorithms (Navarro and Raffinot, 2002; Gusfield, 1997).

Improved programs for similarity search, such as Diamond and MMSeqs, provide sensitivity comparable to BLAST but at much higher speed. MMSeqs achieves this by using a "double match" approach, where it extends k-mers (substrings of nucleotides of length k contained within a sequence) as much as possible without requiring exact matches. Additionally, MMSeqs can speed up searches by distributing queries across multiple processor cores and splitting the target database into smaller parts that can be processed on different servers. DIAMOND in comparison to BLAST uses a reduced alphabet and database blocks to reduce memory footprint. Also, larger and spaced seeds in different shapes are used by the program. Other DIAMOND improvements over BLAST include compressed double indexing of query and reference proteins to reduce main memory accession (data locality). DIAMOND is the fastest program for similarity search available (Figure 5) (Buchfink *et al.*, 2021; Steinegger and Söding, 2017; Buchfink *et al.*, 2014).

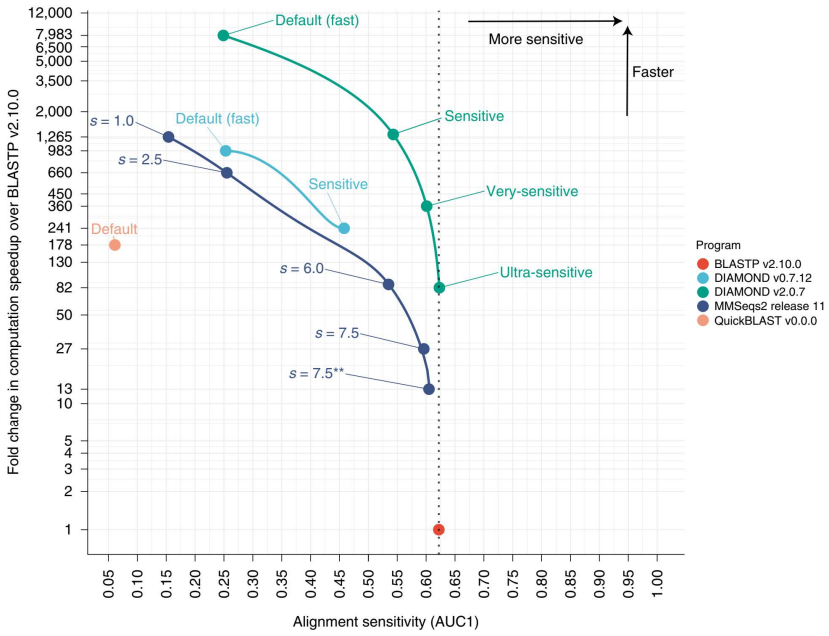

Figure 5. Benchmark of DIAMOND, MMSeqs2 and BLASTP using various sensitivity modes, images taken from Buchfink et al., 2021.

A commonly planted question when trying to find similar patterns in proteins is whether the sequences have total or partial similarities with other sequences (Zhang et al., 1998). Sometimes proteins present similar domains and are statistically found to be homologous. Homology, when found by similarity search, cannot assure every part of the protein has a homolog in the other. Local alignments can be performed to identify the most similar regions between two sequences. These domains are overlooked sometimes by the models present in databases such Pfam or InterPro since homologs can be distant (Pearson, 2013; Gonzalez and Pearson, 2010). Local alignment is made following either Needleman-Wunsch or Smith-Waterman algorithms in which evolutionary matrices play a significant role by measuring the weight of each difference/substitution in the sequence. This similarity on protein domains provide useful information for predicting structure, evolution and function (Chao *et al.*, 2022).

Survival of initial patterns after a second alignment can indicate the conservation of the pattern across sequences. MSA can be a useful tool on detecting conserved motifs (Mansour et al., 2009). This technique extends the local alignment algorithm with different methods to align every sequence in a database (Chao et al., 2022; Barton and Sternberg, 1987). Phylogenetic inference is a result of processing MSAs results, some errors in sequencing, generating ambiguously aligned sequences, and structural annotation can drive to mistakes in these evolutionary reconstructions (Di Franco et al., 2019; Ashkenazy, 2019; Damkliang et al., 2017).

2 Objective

To deepen into our knowledge of the flavoproteome of *M. tuberculosis* by creating an efficient method of searching for orthologs, and predicting potential catalytic activity based on protein sequences.

3 Method

3.1 Experimental design

The aim of the proposed methodology was to provide both an efficient and accurate approach to the quest for information regarding the *M. tuberculosis* flavoproteome. As described previously (section 1.3), there are several approaches for predicting protein functions, being those which can combine genetics, structure, protein interactions and sequence similarity the ones with higher prediction performance (Lin *et al.*, 2024; Sivashankari and Shanmughavel, 2006).

Despite the fact that protein function is a wide term that comprehends cellular location of a protein, molecular pathway, splice forms, regions, protein interaction and stability, most flavoproteins are known to have enzymatic activity. Catalytic activity of flavoenzymes can be well characterized by molecular function and metabolic pathway. Many useful approaches rely heavily on deep learning and/or networks to consider all mentioned protein function characteristics (Lin *et al.*, 2024; Stringer *et al.*, 2023). The complexity of a method that accurately predicts catalytic activity by taking information from previous annotations should not necessarily be high. Previously known information, cross-referenced annotations, improved algorithms for similarity search along with genetic relationships may provide enough information improving efficiency by lowering pipeline complexity.

The main theoretical assumption behind the method used to retrieve valuable data, especially catalytic activity of a particular enzyme, was the already mentioned ortholog conjecture (section 1.4). As is shown below (figure 7), despite the ortholog conjecture is still debated, data provided by spanning archaeal, bacterial and eukaryotic proteomes support the idea that orthologs can be a highly accurate predictor of enzyme functions in the way of enzyme commission numbers (around 95% percent) (Altenhoff *et al.*, 2016).

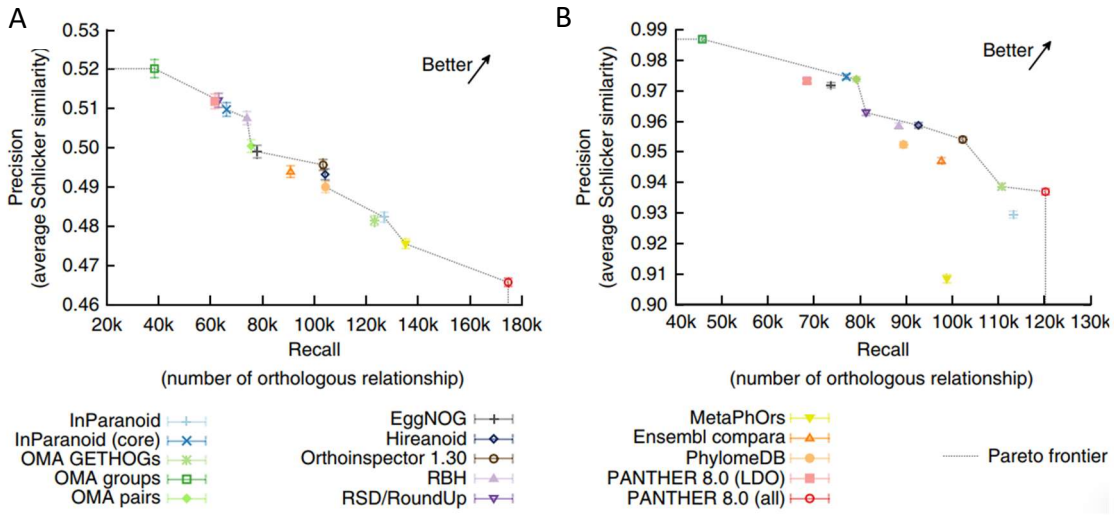

Figure 6. Benchmarking of different ortholog search methods and its predictive precision using Schlicker semantic similarity (Schlicker *et al.*, 2006) *(A) experimentally supported GO annotations and (B) Enzyme Commission (EC) numbers. Error bars indicate 95% confidence intervals, images taken from Altenhoff et al., 2016.*

For gene ontology terms, similarity between prediction and experimental annotation of ortholog search methods is known to be approximately 50%, depending on similarity measure and genetic distance of species. When similarity measure or species genetic closeness changes, this similarity for gene ontology terms varies and indicates sometimes higher similarity between paralogs than

between orthologs. (Stamboulian et al., 2020; Altenhoff et al., 2016; Altenhoff et al., 2012). Because the description of the function of flavoenzymes is centered on catalytic activity, they could accurately be characterized by EC numbers. Ontology terms, sometimes only annotate some functional characteristics that are often related but are not necessarily indicative of similar catalytic activity (Aleksander *et al.*, 2023).

Reciprocal best hits (RBH), also known as bidirectional best hits (BBH) are those pairs of sequences whose similarity is recognizable between them, with no gene sequence in the whole species genome being more similar to the other in its species genome and *vice versa* (Sivashankari and Shanmughavel, 2006). RBHs are usually employed to find orthologs, sometimes used to infer directly orthology by replacing phylogenetic inference, that uses trees along with other genetic considerations. What is found about RBHs is that they highly agree with phylogenetic inference, and that the vast majority of genes from 'syntenic' operons in some archaeal and bacterial genomes are found to be RBHs in closely related species (Dalquen and Dessimoz, 2013; Wolf and Koonin, 2012; Altenhoff et al., 2012).

Nevertheless, RBH have their limitations and although their precision for the goal of this methodology is expected to be high, this is a conservative method, and its formulation leaves room for unmet theoretical considerations. RBH can only detect one-to-one orthology. Therefore,in the presence of a different number of duplications in each species, it doesn't detect all orthologs but the most similar pair. Furthermore, the case of gene loss is not fully contemplated in this approach. Therefore, after such an event occurs, pairs of genes that are caused by duplication and subsequent speciation can be detected as orthologs and *vice versa*, causing FP and FN respectively (cases c and d in figure 8). It is true that a wide percentage of RBH are orthologs, but it is estimated that they are only representative of 40-45% of them (Dalquen and Dessimoz, 2013). RBH performance in terms of FPs and FNs is represented in the figure below (figure 7).

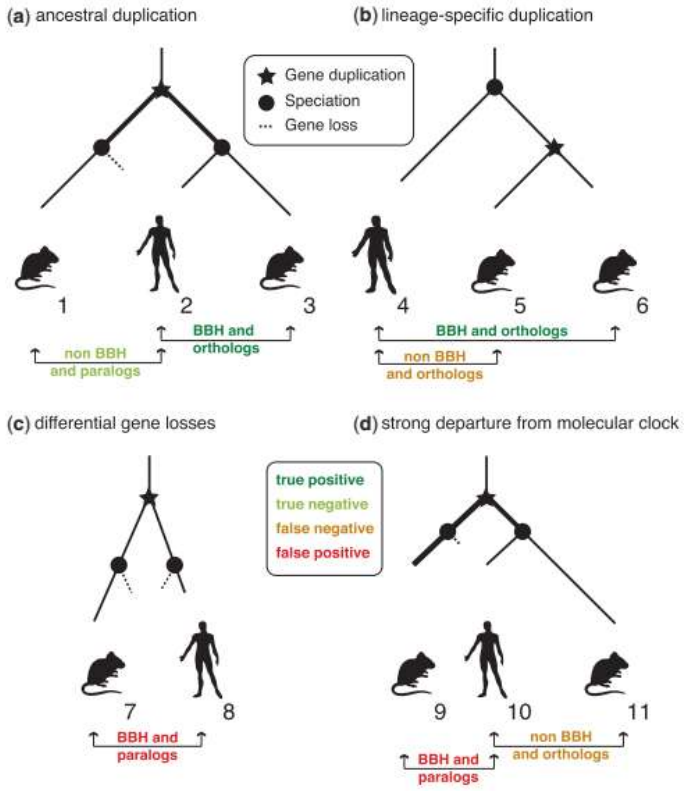

Figure 7. Performance of RBH in conceptual examples, images taken from Dalquen and Dessimoz, 2013.

Unidirectional best hits (UBH) are best hits from each taxonomic species and the result of the very first search that is performed in order to obtain RBH. UBH are more numerous but also more diverse in terms of genetic relationships in comparison to RBH. If RBHs comprehend only so called one-to-one orthologs, UBHs may include gene loss and speciation after duplication orthologs not contemplated by RBH. Unfortunately, UBH might also mislead more paralogs as orthologs. UBHs as well as RBHs are restricted to the best hits that are annotated in the database, not the actual best hits of the whole proteome of each species. Moreover, non-reciprocity in UBH causes lower precision in ortholog detection since these best hits are not again compared against query taxonomic species annotated proteomes, thus increasing this annotation inequality bias.

With the aim of developing a method that could efficiently analyze the whole identified proteome of *Mycobacterium tuberculosis,* a RBH search was performed using the Diamond blastp and UniProt database for the 184 protein queries. A custom flavoprotein database with catalytic activity annotations was the reference for the first RBH search. The idea of creating a custom database arose from the need for efficiency, as Diamond requires database format conversion and does not have servers to perform the search. Moreover, the number of sequences and their characters limits the speed of the search algorithm and data handling.

The custom database was designed to include only valuable data trying not to compromise output bias. Thus, sequence data selection was performed considering protein cofactor nature and activity annotation availability, in this case, reference proteins are only flavoproteins with catalytic activity related records. For the RBH and UBH the flavoprotein database filter used to query against may not make much difference in predictive precision since homology is most likely to be only significant within flavoproteins. The amount of orthologs retrieved would be restricted by the number of proteins present in the database, this can be mainly due to catalytic activity annotation presence. While database bias could be less significant to RBH, UBH could be more affected by filtered database bias due to non-reciprocity and higher annotation inequality bias.

To assess protein function, predictive precision recall and annotation coverage of the method parameters were calculated for previously characterized proteins (queries). For both RBH and UBH calculated parameters were represented for every protein to detect possible tendencies related to bias and precision in each case.

3.2 Workflow

The method implies two similarity searches against different databases. For this task, query sequences are obtained from NCBI's database and then, copied to a fasta file that contains the *Mycobacterium tuberculosis* known flavoproteome using Entrez and **SeqIO** modules from **Biopython**. The database of reference proteins is downloaded from UniProt database writing queries in the way of key words FAD, FMN and F420, and filtering UniProt website output selecting only those proteins whose catalytic activity is annotated. This custom database has to be formatted to be Diamond's blastp input by using **makedb diamond's command** that converts fasta files into diamond files (figure 8).

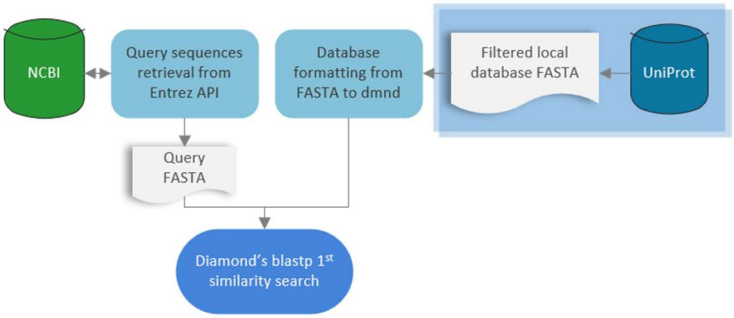

Figure 8. Workflow steps leading up to the first similarity search.

The second similarity search database consists of sequences, descriptions and taxonomy (scientific names) from the first search unidirectional best hits homologs. For database construction scientific names of the species are easily extracted from the filtered fasta database file and UBHs are obtained by selecting homologs with the minimum E-value for each taxonomic species and filtering by query and subject alignment coverage criteria. Coverage filtering criteria selection is performed and optimized with parallel computing for data frame transformations with **dask** module. Proteins that were queries in the first search are the ones that need to be compared with the output's best hits of this second search to find out which of them are reciprocal best hits. To match these identities, first search queries' codes are translated from NCBI code to UniProt code with **UniProt ID mapping tool** via **unipressed** python module.

EC numbers, pathway, description and other annotations from UBH are then obtained, the protein description is retrieved from the database fasta file, while other annotations are retrieved from **UniProt API** (programming interface for UniProt search). This process requires connecting the database for each ID and accessing concrete information through its hierarchized schema, accession optimization was achieved by an asynchronous execution module named **concurrent.futures**. Once the best hits are selected, protein sequences are obtained from the first database fasta file. Reciprocal best hit 2nd search is then performed.

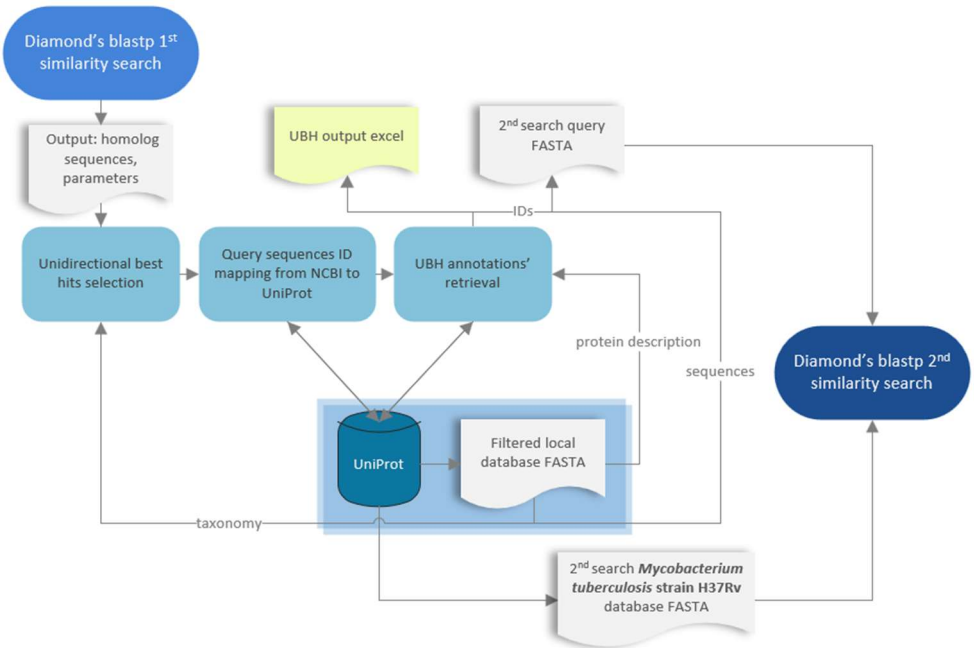

Figure 9. Simplified steps of the workflow leading up to the second similarity search.

Once the second search is finished, hits are selected from homologs with the same criteria as the first search homologs (filtering by query and subject coverage criteria). Both UBH and all filtered hits resulting from the second search are merged into a table containing alignment parameters and annotations from the first search. There, only the UBHs that are reciprocal in the second search filtered hits are present, that is only RBHs (Figure 10).

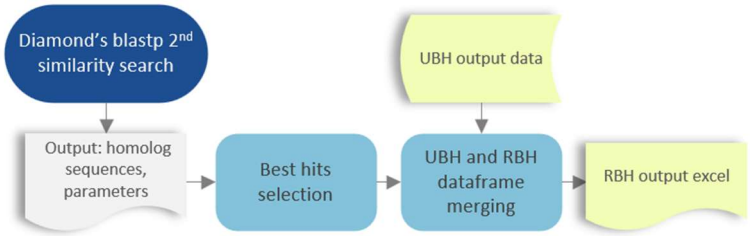

Figure 10. RBH output refinement from the second similarity search.

3.3 Parameters

The lowest E-value was the criterion for selecting the best hits for each taxonomic species, with a threshold of 40% was established for query coverage and subject sequence coverage in the alignment. By default, the maximum E-value threshold was set to 0.001. The number of maximum hits for each query was selected to be higher than the expected hits to obtain all significantly similar sequence alignments. The substitution matrix used for the alignment was the default (BLOSUM62), as it detects more hits. The Diamond's ultra-sensitive mode was used to obtain a sensitivity similar to BLAST's.

3.4 Programming languages

Python 3.12.7 and bash (Ubuntu 24.04.1) were used for the main script in which figure 9, 10 and 11 workflow steps are performed. Python was used to build the script for finding orthologs as well as to merge tables of previous data. In practice, an ortholog search script was employed to extract taxonomy from similarity search hits and to map protein identity numbers between databases, as well as to integrate diamond commands into the Linux terminal for similarity search and to retrieve and parse fasta files for databases and input query files.

Biopython with the Entrez and **SeqIO** modules were employed for generating query fasta files, and the **unipressed** (UniProt API) module was used to retrieve information from the Uniprot website and map IDs between databases. The modules **dask** for parallel computing and **concurrent.futures** for asynchronous execution modules were used to enhance computational efficiency. Other default packages were employed to transform and read data such as Pandas and Regex. Communication between the Linux terminal and the python script was facilitated by Subprocess module. Bash was the language that allowed fasta file combination as well as diamond installation and running.

For the statistical analysis of results and graphical representations, modules such as Pandas, NumPy, Matplotlib, and Seaborn from Python 3.12.7 were used.

3.5 Environment and programming tools

Visual Studio Code was the environment of choice to build Jupyter notebook files and scripts. WSL Linux (Ubuntu 24.04.1, kernel: GNU/Linux 5.15.167.4-microsoft-standard-WSL2 aarch64) distribution on Windows was the environment that allowed diamond installation and execution as well as fasta file handling.

3.6 Databases

NCBI and UniProt were both used for similarity search; NCBI was the database query sequencies were retrieved from and UniProt was the database for reference protein sequences. Already annotated NCBI codes of *Mycobacterium tuberculosis* proteins were the input for the first similarity search, so NCBI was the database employed to obtain query sequences. UniProtKB was the database of choice to query, since it is highly cross-referenced, has a large number of manual and computational annotations from several databases, and can be easily filtered and accessed programmatically (Bateman *et al.*, 2023).

3.7 Bioinformatic programs

Diamond v 2.1.8-2 for Linux (diamond-aligner), diamond blastp algorithm using ultra-sensitive mode enabled homologs' search.

3.8 Equipment

The equipment used was a laptop with the following characteristics: Snapdragon (TM) 7c @ 2.40 GHz 2.40 GHz ARM based processor, 8.00 GB (7.57 GB usable) RAM.

4 Results and discussion

4.1 Functionally uncharacterized flavoproteins

The partially annotated *M. tuberculosis* H37Rv strain flavoproteome was known to have reviewed enzymatic functions for some proteins, so is to say, cross-reviewed EC numbers and metabolic pathways from Mycobrowser and KEGG were found (Montesa et al., 2023 TFG). However, a large number of the identified flavoproteins lack a clear function, being therefore unlikely to be suitable for biomedical or biotechnological purposes. Of the 184 flavoproteins listed for *M. tuberculosis* so far, 133 have incomplete functional annotations.

4.2 Method performance

The script created for RBH ran in 48 min and 27 s and queried the 184 identified *M. tuberculosis* H37Rv strain flavoproteins. In the first similarity search 178 of all the queries were aligned with other 1110383 high scoring pairs (HSPs) in 17 min and 13 s. Annotations for the resulting 30260 UBH were retrieved in 13 min and 44 s and were searched and aligned against the *M. tuberculosis* H37Rv strain proteome obtaining 495813 HSPs in 4 min and 25 s. Annotations for the resulting 12676 RBHs were retrieved in 4 min and 1 s. Other processes described in the method above (section 3.2) took around 10 min to complete.

Of the initial similarity search, 178 proteins were aligned. Within this set, 117 queries met the criteria for query alignment coverage remaining as UBH queries, while 49 of them were reciprocal and met the coverage and E-value criterion in the second search. The number of orthologs identified for each protein is not uniform among queries, as shown in the bar plots in figure 11 and showing inherent sequence and functional annotation bias of the database.

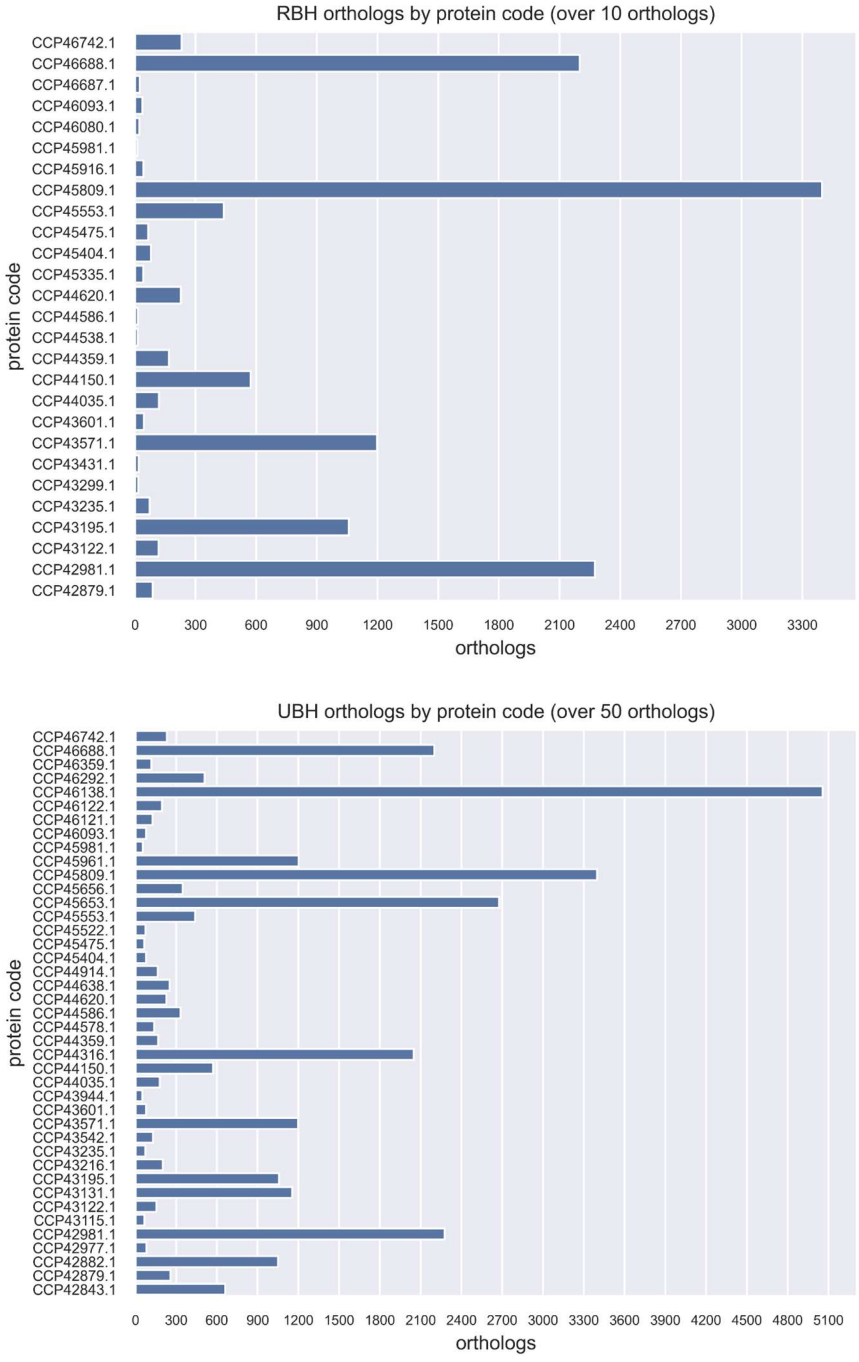

Figure 11. Number of orthologs obtained for each query by RBH and UBH for queries that have a minimum number of orthologs of 10 and 50.

Of the 30260 UBHs and 12676 reciprocal best hits, the orthologs with the corresponding annotations were the ones considered for assessing precision, recall, count and annotation presence. Annotations obtained were EC numbers, pathways, protein description, cofactors and

reactions. As enzymes can possess different catalytic functions, up to two different EC number annotations, pathways and reactions were obtained for each protein, as demonstrated in previous annotations.

4.2.1 Prediction vs. reference annotations evaluation parameters.

For assessing predictive precision and recall of the method, previous annotations regarding ortholog's EC numbers and pathways were considered (Montesa et al., 2023 TFG). Note that these annotations are computational and experimental, and not all are based on direct evidence.

Whether for pathway or EC number annotations, true positives (TP) were calculated as the number of orthologs for each protein that contained the first or second pathway or EC number that this protein showed in previous annotations. False positives (FP) were those orthologs of the protein that showed one or two of these EC numbers or pathways different from the ones recorded. False negatives (FN) were estimated as orthologs with no EC number or pathway annotation, despite the fact that the query's EC number/s or pathway/s had been previously described. True negatives cannot be defined, since non-existent annotations of an ortholog cannot be considered false or true. Therefore, they are included in the FN estimation.

$$TP = number\ of\ orthologs\ matching\ one\ or\ two\ annotations\ with\ its\ query\ annotation$$

$$FP = number\ of\ orthologs\ all\ of\ whose\ annotations\ do\ not\ match\ the\ query\ annotation$$

$$FN = number\ of\ orthologs\ \ of\ an\ annotated\ query\ with\ no\ annotations$$

Precision is defined as the number of TP obtained from the set of annotated orthologs and queries. Recall represents the number of TPs from annotated queries and from not annotated and unnnotated orthologs corresponding to previously annotated proteins in terms of pathway or EC number. To assess the relative quantity of annotations (annotation coverage) of hits in the UniProt first searched custom database, true positives and FP divided by the number of orthologs were used as an estimate for both EC numbers and pathways. Note that EC numbers' and pathways' recall and precision are calculated separately, becaise orthologs do not always have both annotations. They are semi-automated and computationally annotated in the database, thus the EC number and pathway annotations are not always concordant of each other in UniProtKB (Bateman *et al.*, 2023).

$$precision = \frac{TP}{FP + TP}$$

$$recall = \frac{TP}{FN + TP}$$

$$ortholog\ number = FP + TP + FN$$

$$annotation\ coverage = \frac{TP + FP}{FP + TP + FN}$$

Note that the parameters described take values between 0 and 1, maximum value responds to self-division and minimum to TP or TP plus FP equal to 0.

4.2.2 Overall predictive precision, recall and ortholog annotation coverage.

General precision, recall and annotation coverage were calculated for all the ortholog counts of all the queries. Figure 12 represents the general parameters as the sum of all orthologs' predictions respect to all corresponding queries annotations.

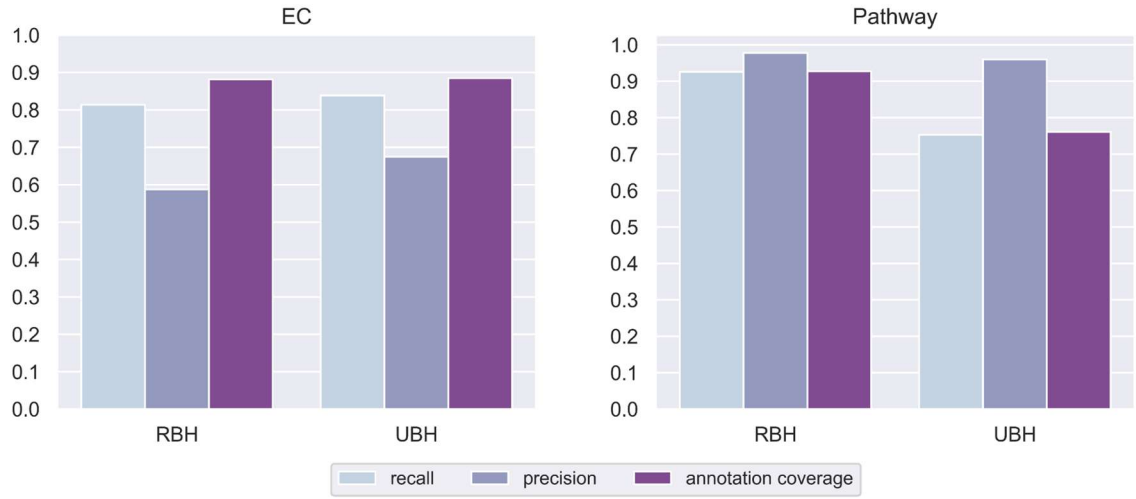

Figure 12. Overall recall, precision and annotation coverage obtained from RBHs and UBHs for the sum of all queries orthologs' annotations regarding EC numbers and pathways.

The general recall and annotation coverage for both RBH and UBH in the case of EC numbers is high, so the method retrieves a high number of well predicted EC numbers from all of the relevant entries and several annotations. EC number predictive precision is lower; there are numerous FPs. Nevertheless, precision is still more than half in both cases and generally higher for unidirectional best hits.

The overall accuracy and recall of pathway prediction is really high, meaning that the number of TP is considerable and FP and FN are less present. FN are more numerous than FP and this difference is greater in the case of UBH methods where FN are even more present (lower recall). That is, the retrieval of actually predicted cases over representative ones is lower than the proportion of correctly predicted pathways over annotated ones. Pathway annotations are generally fewer than EC number annotations for both methods as the database filter is the presence of catalytic activity annotations.

4.2.3 Query predictive precision, recall and ortholog annotation coverage.

For a better understanding of these resulting parameters for both RBH and UBH methods, a representation of precision and annotation coverage of each query was performed (figure 13)

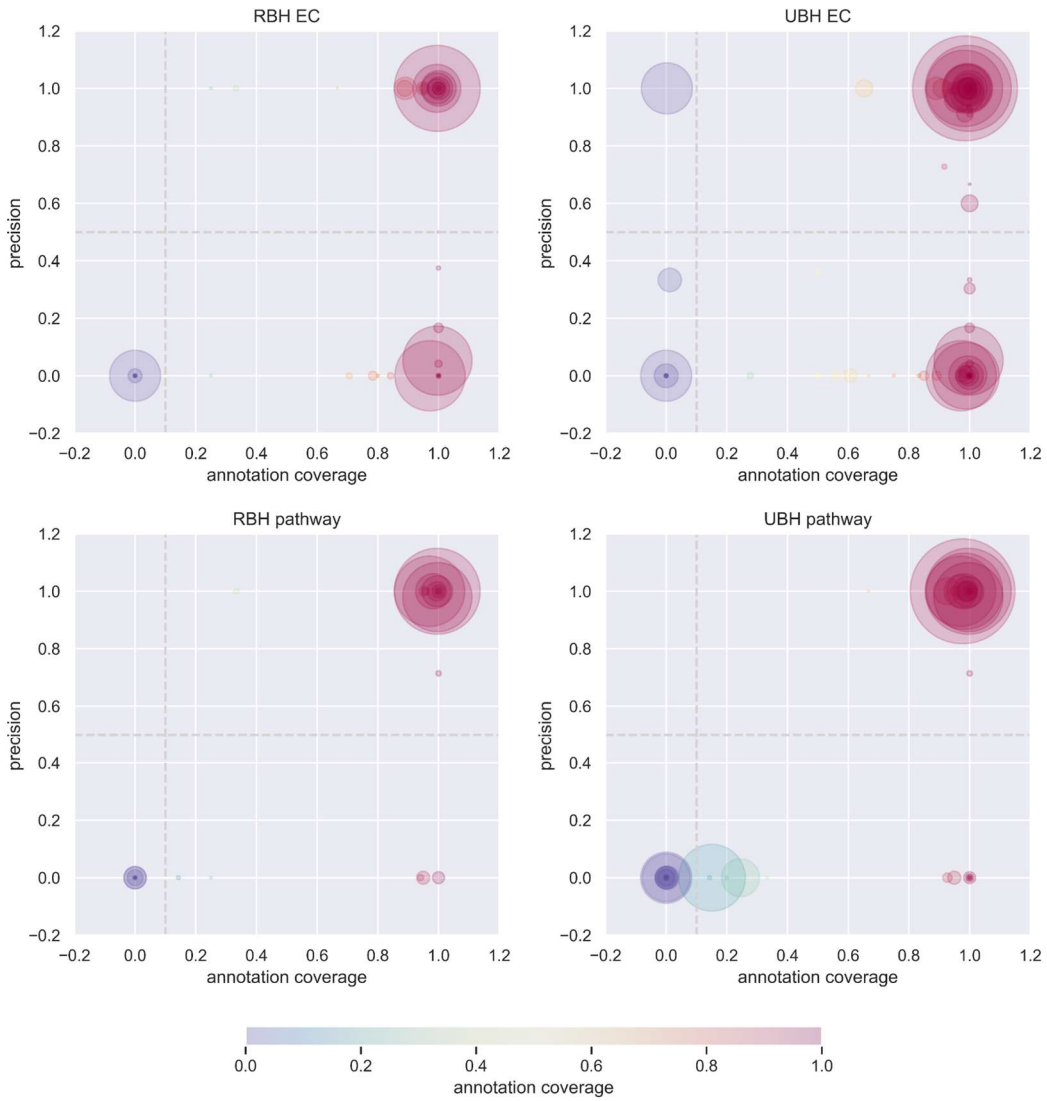

Figure 13. Precision against annotation coverage for each query represented in a scatter plot where the area of the points is proportional to the number of orthologs found for each query and colour enhances annotation coverage differences. Dashed lines represent value thresholds for precision and annotation coverage classifications.

The distribution of EC number annotation queries forms a square, with three corners that are heavily populated, and a few scattered exceptions along the edges. Using thresholds of 0.5 for precision and 0.1 for annotation coverage, proteins are classified as annotated and at least moderately precise. Since most queries are well-annotated, many fall into the annotated category. Among those, a majority show high precision. The UBH dataset clearly includes more queries and orthologs per query, as reflected by its larger, more numerous points on the plot compared to RBH.

For pathway predictions, the data mostly clusters into two corners of the same square setup—one representing high precision and coverage, the other representing low values for both. Compared to EC number annotations, fewer pathway predictions fall into the low-precision group, and those queries tend to involve fewer orthologs. Pathway predictions generally show more true positives and fewer false positives. The same thresholds are applied here: 0.1 for annotation coverage and 0.5 for precision. Similar to EC numbers, the UBH results in more queries being matched, with more orthologs per query. Overall, the most common scenario across both types of annotations

is having high precision and annotation coverage. This tends to correspond with having more orthologs per query.

Precision against recall was also represented in the following figures, figure 14 below represents RBH and UBH results for EC number annotations.

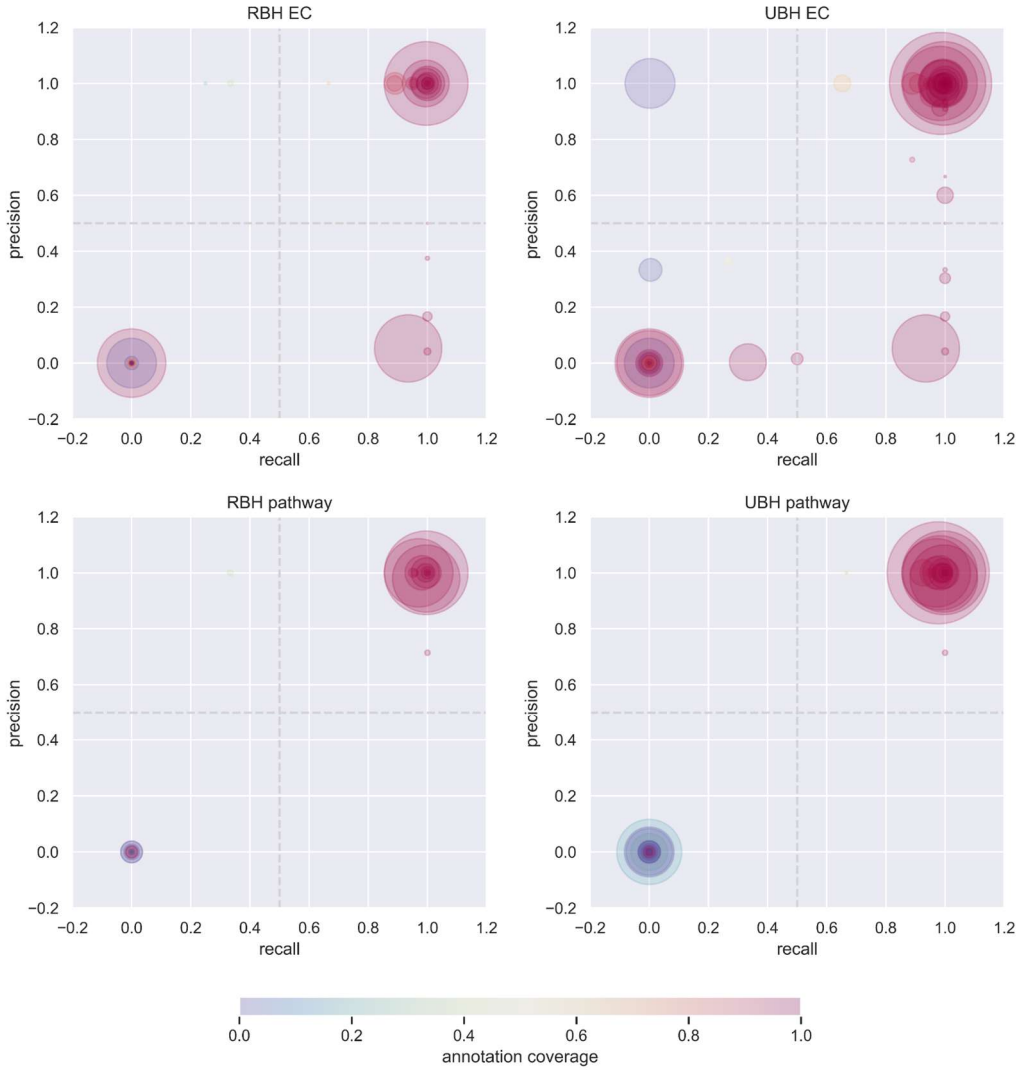

Figure 14. Precision against recall for each query represented in a scatter plot where the area of the points is proportional to the number of orthologs found for each query and colour indicates annotation coverage values. Dashed lines represent value thresholds for precision and recall classifications.

Distribution of results for recall against precision regarding EC number annotations follows a similar pattern as the one described on figure 13 with the exception that points representing very low recall and precision are more numerous. This is because positive values of recall require the presence TPs while annotation coverage positive values can be achieved also through FP presence. In this case thresholds for query results' classification were set on the retrieval of a moderate number of TPs and moderate precision, that is, 0.5 for both precision and recall. High precision and recall classification is populated in both graphics and exceptions to this classification for the query and the contrary case (low recall and precision) are less than the ones described for the classification of precision versus annotation coverage since high values of recall are harder to achieve compared to higher values of annotation. Again, an increase on ortholog number retrieval

is seen for UBH respect to RBH as values of ortholog number and queries paired was shown at the beginning of this section (4.3) and describing efficiency (4.2).

Distribution of results for pathway annotations is polarized in two categories, very low recall and precision and very high recall and precision. Thresholds are set in 0.5 for both precision and recall as for EC numbers. This means that in cases when good retrieval of TPs from significant results is made, TPs are in comparison to FPs very numerous. In both graphics can be also appreciated the increase in the number of orthologs per query and the number of queries paired for UBH respect to RBH.

To better visualize the points that fall under the thresholds 3D plots are made where annotation coverage, recall and precision. Figure 16 represents these parameters for EC number and pathway predictions for each query.

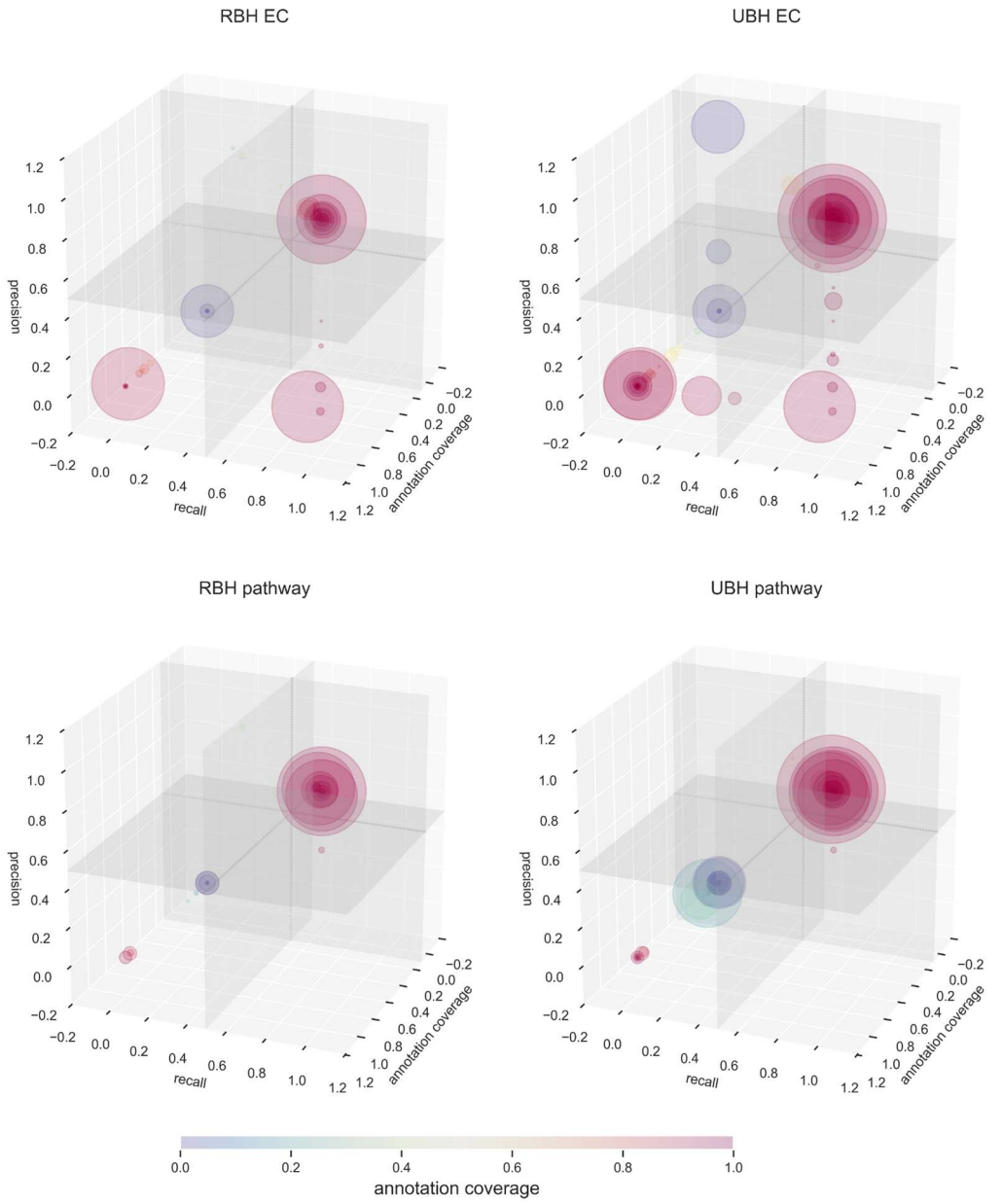

Figure 15. Precision against recall scatter plot for each query, size of the point represents the number of orthologs found for the query and colour enhances annotation coverage values present in the third axis.

Figures 13 to 15 show the distribution seen above in where queries often present extreme values with some exceptions. Classifications in which queries fall are those regarding all annotation coverage, recall and precision thresholds:

- Queries with low annotation coverage, recall, and precision are those with no annotated hits (FNs), due to the stringent annotation threshold that excludes all unannotated orthologs.

- Low predictive precision, high annotation coverage and high recall mean numerous FPs and TP presence.

- Flavoproteins with high ortholog predictive precision, annotation coverage and recall are those with a high number of TPs and a low count of FN and FP.

- High precision, low recall and low or high annotation coverage means a strong presence of not annotated proteins (FNs) with a low count of TPs, that is, poorly annotated orthologs coincide with previous annotations.

- Predictions for queries with low recall and precision and high annotation coverage have FNs and high FP count.

To estimate the number of queries that retrieve valuable information the number of queries for each classification is represented in the figure 16 to estimate the number of queries that retrieve valuable information. Queries with low annotation coverage, recall, and precision are not included in the charts since they do not retrieve annotations.

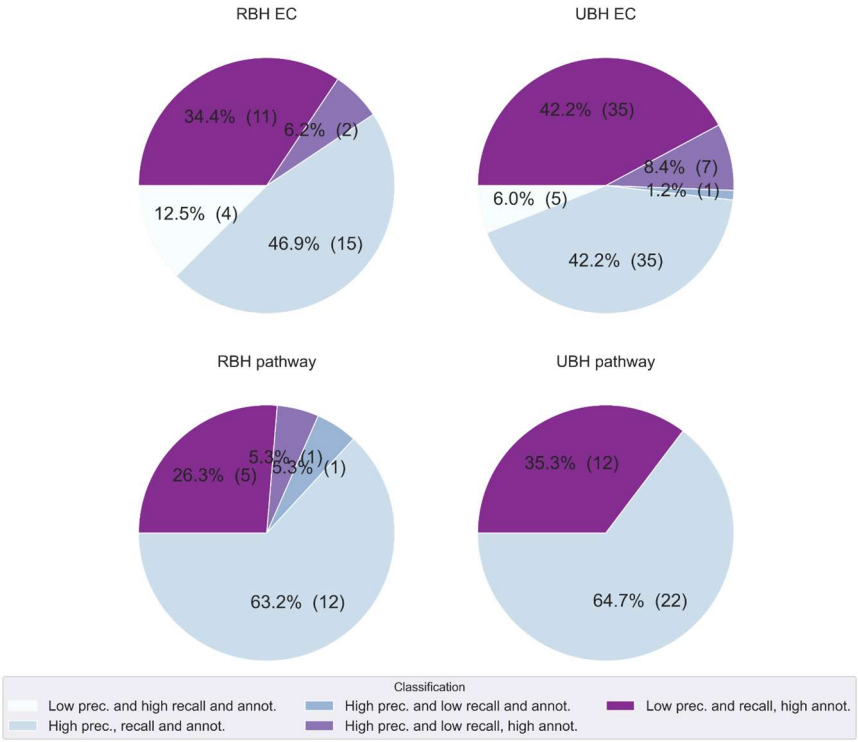

Figure 16. Chart representing the number of queries in each classification for EC number parameters and pathway parameters in RBH and UBH. Classifications are represented by different colours and indicated by its corresponding legend, chart title indicates if parameters are calculated on pathway or EC number annotations and whether they respond to RBH or UBH predictions.

Charts shown in figure 16 show a clear difference between the number of queries in each classification for UBH and RBH for different annotations. The difference in protein number between classifications is higher for RBH where FPs are predominant (purple), having less query percentage than the one where TPs are more numerous (light blue and light purple). This means that the genetic relationships between homologs and queries for UBH produce less predictions in which more than half of the annotated homologs show a correctly predicted pathway or EC number.

For the EC number statistics classifications, white and cyan classifications (figure 16) are those with low count of TPs. The difference between UBH and RBH regarding these classifications is that queries in RBH fall under white classification more often than those corresponding to UBH. While one UBH query falls under cyan classification, RBH do not pair any query that may fall there. Thus, UBH respect to RBH responds to a lower number of queries that have a low count of TP, although one of them shows low annotation coverage and high predictive precision.

Making a closer look to white classification queries' statistics, annotations indicate that EC numbers constituting FPs are similar to those previously recorded.

Table 1. White classification EC number UBH statistics and annotations for white classification in figure 20, including protein NCBI code, description, pathway and EC number of each query respectively for first columns. The last three columns show the EC number annotations respect to each query's hits and the number of hits (counts) containing the annotations shown in the previous two columns (EC 1 and EC 2). EC annotations for homologs are divided in columns EC 1 and EC 2, since two EC numbers are retrieved, when available for each homolog.

Protein code	description	Pathway	precision	annotation	EC	EC 1	EC 2	counts
CCP43138.1	F420-dependent glucose-6-phosphate dehydrogenase	Carbohydrate metabolism	0.333333	1.000000	1.1.98.2	1.1.98.2		3
						1.5.98.2		6
CCP45916.1	NADPH-ferredoxin reductase FprA		0.166667	1.000000	1.18.1.2	1.18.1.2		7
						1.18.1.6		35
CCP45981.1	Flavin-containing monoamine oxidase AofH		0.303571	1.000000	1.4.3.4	1.4.3.2		39
						1.4.3.4		17
CCP46687.1	Glutamate synthase [NADPH] small chain	Amino-acid biosynthesis	0.041667	1.000000	1.4.1.13	1.3.1.2		1
						1.4.1.14	1.4.1.13	1
								21
						1.8.1.19	1.18.1.2	1
CCP46688.1	Glutamate synthase [NADPH] large chain	L-glutamate biosynthesis	0.052464	0.996364	1.4.1.13	1.4.1.13		47
						1.4.1.14	1.4.1.13	68
								2068
						1.4.7.1		5
						2.7.1.148		1
						2.7.13.3		1
						3.4.13.22		1
						5.6.2.2		1
								8

Tables 2 and 3 suggest that in this classification EC numbers have low precision, but TPs are still high in number. Also, FPs in this category may represent another possible function for the query that is present in other homologs. As shown in figure 16 for purple classification queries often present null precision and recall, since many predictions are FPs and there is no TP presence and retrieve often no valuable information.

For EC number predictions RBH similarity yields high precision and annotation coverage for 53.1% of queries, valuable in 12.5%, and low value in 34.4%. UBH similarity shows high precision and annotation coverage for 51.8% of queries, valuable in 6%, and low value information in 42.2%. For

pathway predictions RBH achieves high precision in 72.3% of cases, with 27.8% being of low value. UBH has 64.7% high-precision predictions and 5.3% low-value results. Overall, UBH tends to yield fewer queries with valuable annotations compared to RBH. Pathway predictions generally show a higher proportion of valuable results than EC number predictions.

As discussed in section 3.1, the broader genetic diversity of UBHs often reduces prediction accuracy, whereas RBHs and certain closely related paralogs usually lead to more accurate results. While more genetically diverse orthologs can sometimes improve predictions, an increased number of orthologs doesn't necessarily guarantee higher precision. This explains the difference between general precision rates and the proportion of valuable predictions per query, as seen in the comparison between figures 12 and 16. Precision often benefits from having multiple orthologs, but this is not always the case, as shown in figures 13 to 15.

To understand how the alignment of queries with their orthologs differs between classifications, boxplots were created for each combination of annotations (EC numbers or pathways) and orthology (UBH or RBH), showing the mean E-value for each query.

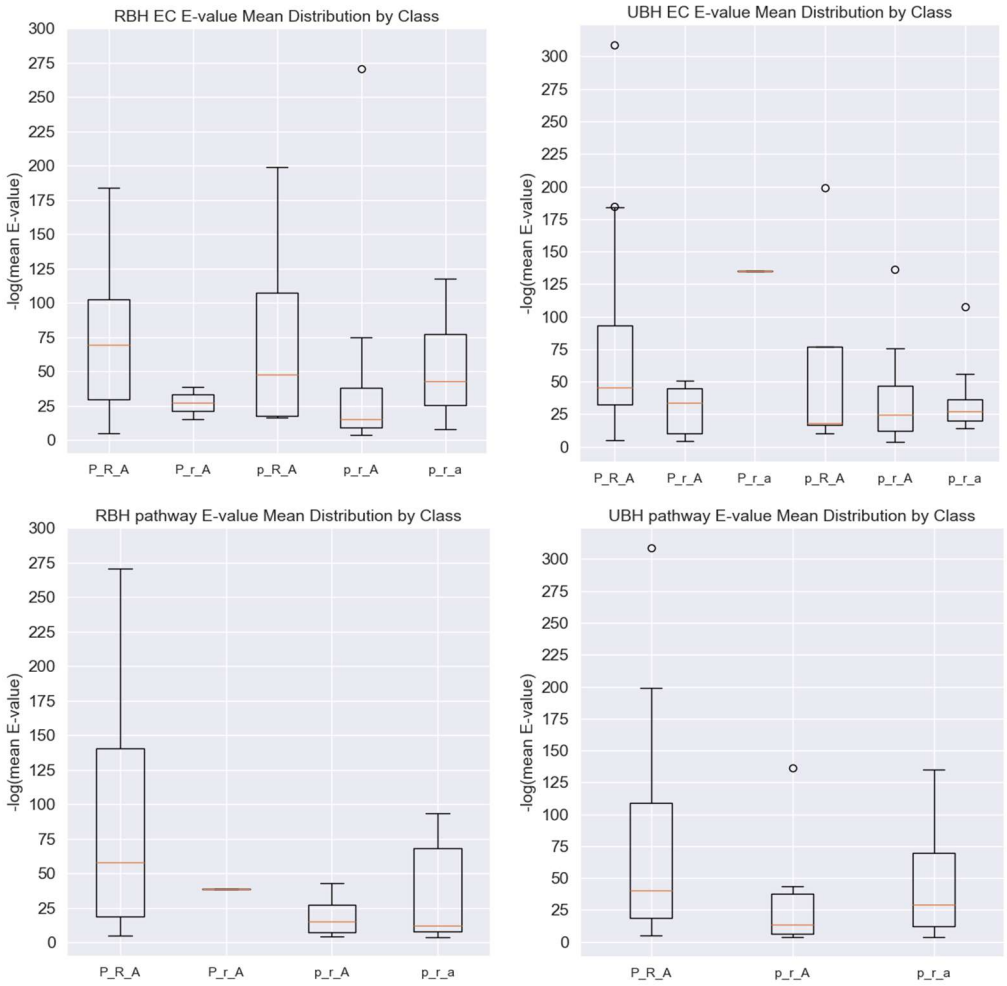

Figure 17. Negative base-10 logarithm of E-value means for each query for the different classifications made. Each graphic title indicates the nature of the data, coming from RBH or UBH or comparing pathway results or EC number results. Also, classifications are named in the x axis with a combination of capital and small letters referring to the first letter of the parameter's; precision, recall and annotation coverage. Capital letters indicate parameter threshold overrides, while lowercase letters represent statistics below the threshold.

Generally, figure 17 shows differences in classification. High precision and recall classification show higher -$\log_{10}$(mean E-value) mean. The other classifications vary since many of them still present TPs or FNs that could be significant alignments but are not annotated. The only classification with a high number of FPs is the one with low precision and recall and high annotation where -$\log_{10}$(mean E-value) mean is equal to or less than 25 in all cases. Thus, an E-value threshold of approximately 1E-30 is thus a good criterion for determining similarity significance within different orthologs for the same query. Note that the low precision, annotation coverage, and recall classifications are not represented in figure 16, but are represented in figure 17.

To quantify the precision of each query for a given protein, a representation of EC number annotations precision for each query is indexed with the protein description (figures 18 and 19). Only queries with homologs presenting EC number annotations are included in the plot.

Figure 18. EC number precision for queries described. RBH results are represented in blue and UBH in red, common queries with both RBH and UBH are represented in a different plot overlapping colour codes and bar size.

For Acyl-CoA dehydrogenase enzyme FadE variants, low precision is observed in the prediction of catalytic activity. This is due to the large number of similar gene variants with similar activities on different substrates. UBH predictions are proportionally more inaccurate than RBH predictions, but there are significantly more total UBH predictions.

RBH predictions whose queries are also found in UBH ones generally show enhanced precision. It is also found in UBH predictions an increase in precision respect to RBH ones. These increase in the precision of some null precision RBH predictions it is most likely due to the increase in the homolog number and maybe to the heightened diversity in UBH.

A representation of pathway precision is indexed with protein description in figure 19 to show precision on pathway annotations for each query.

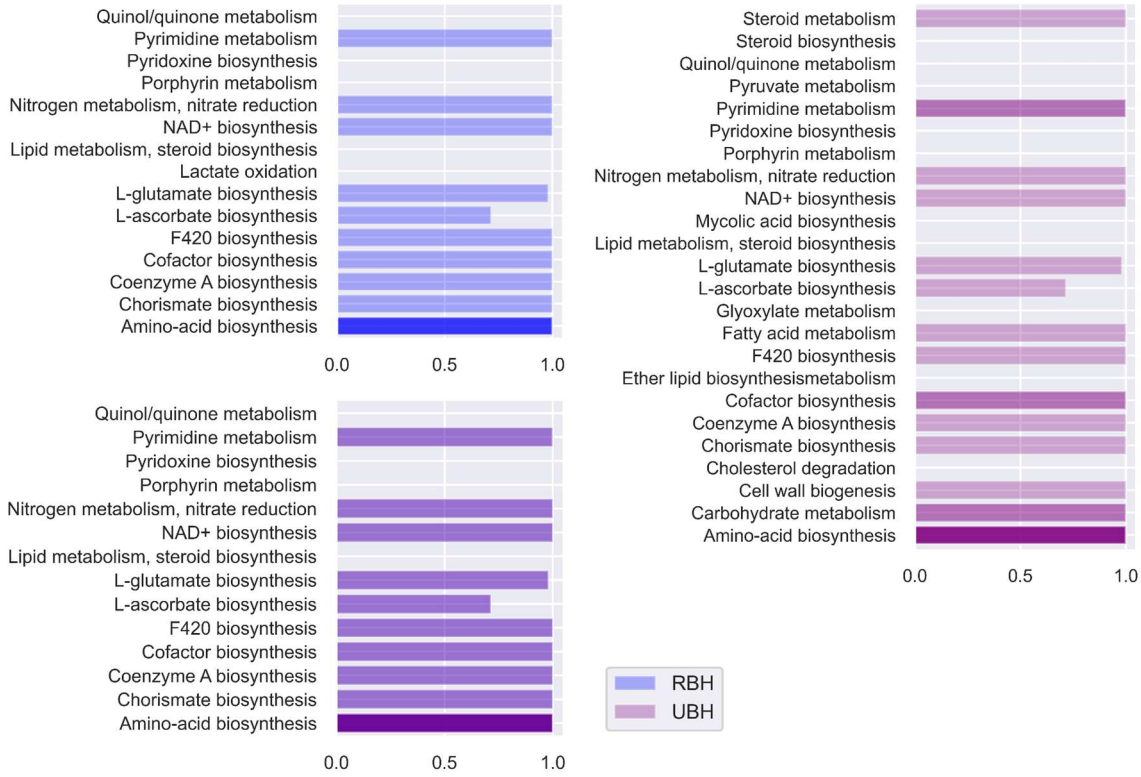

Figure 19. Pathway precision for queries described. RBH results are represented in blue and UBH results in red in different plots. Common queries with both RBH and UBH are represented in a different plot overlapping colour codes and bar size. Colour intensity indicates the number of equally named pathways.

UBH respect to RBH shows an increase in the number of queries with no predictive precision for proteins that are part of pathways regarding lipid metabolism, steroid biosynthesis or both. Predictions for pathway in both RBH and UBH coincide in precision. If hits for common RBH and UBH queries are increasing in number, they maintain precision value in UBH.

Generally, low predictive precision may be due to scarcity of annotations/sequence in the filtered database and/or low predictability of the genetic relationship mix in some cases. A comparatively high number of orthologs for the combination of annotations is often a predictor for correctness. A low or null predictive precision is shown generally shown for lipid metabolism and/or steroid biosynthesis pathways and for exact catalytic function of Acyl-CoA dehydrogenase enzyme. Alignment significance (E-value) plays a crucial role on determining protein similarity even when previous annotations are not coherent with orthologs found. As was described in section 4.2.1 previous annotations found can be automatic and not fully precise despite it is useful to rely on

them and coherence between orthologs' annotations such as reaction and reaction code (RHEA), cofactors, description, pathway and EC number.

4.3 Biological interpretation of RBH unknown protein results

Of the 49 queries with orthologs found in RBH, 5 queries present orthologs with new information only referring to EC numbers, 3 present new information in ortholog's annotations only respect to pathways and 3 respect to pathways and EC numbers. Some of the hits found for the queries have low significance for the alignments and therefore no predictability. Some of the information found outwrites previous annotations and better describes the function of the query. Tables 3-7 describe predictions and known information for a few orthologs that are functionally diverse or insignificant (E-value > 1E-30). The remaining tables describe the most significant orthologs and their majoritarian functions, which are found in the annexes.

4.3.1 Predictions for queries with new information only regarding EC numbers:

Table 2. Orthologs found for the indicated query. The top panel identifies query code and previous annotations listing any identified pathway, Mycobrowser category and EC number. The bottom panel indicates ortholog's features retrieved from uniport and refer to hit description, up to two pathways and up to two EC numbers. The final column counts how many of these orthologs share the same characteristics. Tables 3-7 share this structure and refer to queries with lack of catalytic activity information, that is, uncomplete or inexistent EC number inference or evidence that may be completed with RBH predictions.

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP42857.1	F420-dependent hydroxymycolic acid dehydrogenase	Mycolic acid biosynthesis	Intermediary metabolism and respiration	1.1.98.-

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
5,10-methylenetetrahydromethanopterin reductase		1.5.98.2			4

Despite significant protein similarity being found for four orthologs with a function involved in methanogenesis, methanogenesis from CO_2 is not part of the pathways for *M. tuberculosis*. 5,10-methylenetetrahydromethanopterin reductase has a completely different function from the previously inferred for CCP42857.1 despite both involve the same riboflavin cofactor F420 and participate in redox processes (https://www.kegg.jp/pathway/mtu01200). E-value for orthologs in comparison with other RBH found is high, varying from 2E-12 to 7E-27, and protein similarity from 29% to 37%. Thus, RBH in this database cannot infer a more precise function for this query than the one previously searched.

Table 3

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP43226.1	GMC-type oxidoreductase		Intermediary metabolism and respiration	1.1.-.-

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
Long-chain-alcohol oxidase		1.1.3.20			1

The name that was given by previous annotations for this protein is very general and it is coherent with what is found about the only ortholog retrieved. The function described for EC number of the ortholog suggest that this protein oxidizes long-chain fatty alcohols and its best substrate is

dodecyl alcohol. The E-value that came out of the search homolog is low (2.52E-37) so similarity is significant.

Table 4

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP43640.1	NAD(P)/FAD-dependent oxidoreductase		Intermediary metabolism and respiration	1.14.13.-

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
Flavin-containing monooxygenase		1.14.13.8		1.6.3.1	1

While the previous knowledge inferred for the query does not define a precise catalytic function. The only ortholog found for this protein has a first EC number function that matches the previous annotations but does not take part in *M. tuberculosis* metabolism. The ortholog was found in an evolutionarily distant species of spider called *Pardosa pseudoannulata* and the HSPs was found to have a low identity percentage (21.9%) and a relatively high E-value (1.61E-10).

Table 5.

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP44538.1	L-gulono-1,4-lactone dehydrogenase	L-ascorbate biosynthesis	Intermediary metabolism and respiration	1.1.2.-

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
Long-chain alcohol oxidase	Lipid metabolism; fatty acid metabolism	1.1.3.20		1.1.3.20	1
D-arabinono-1,4-lactone oxidase	Cofactor biosynthesis; D-erythroascorbate biosynthesis; dehydro-D-arabinono-1,4-lactone from D-arabinose: step 2/2	1.1.3.37			3
L-gulonolactone oxidase	Cofactor biosynthesis; L-ascorbate biosynthesis via UDP-alpha-D-glucuronate pathway; L-ascorbate from UDP-alpha-D-glucuronate: step 4/4	1.1.3.8			10

Similarly to the function described for the query the ortholog L-gulonolactone oxidase was found to be the more numerous and the most significantly similar with an E-value up to 2.72E-73.

Table 6

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP45062.1	FAD-linked oxidoreductase		Intermediary metabolism and respiration	1.-.-.-

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
D-lactate dehydrogenase		1.1.99.6			1

Despite the description for the query protein is very general, it is coherent the only ortholog that came out of the search has a high significance with an E-value of 2.19E-88 and 38.1% identity. The enzyme's EC number can then be predicted as 1.1.99.6, this function is involved in pyruvate metabolism pathways.

Table 7. Summary table for queries with lack of catalytic activity information inference or evidence that are completed with RBH predictions. Indicated information includes Query code, description, pathway and EC, further predictions are

made in green. Predictions that indicated a more accurate function are taken in account for changing corresponding description.

Query code	Description	Pathway	EC
CCP42857.1	F420-dependent hydroxymycolic acid dehydrogenase	Mycolic acid biosynthesis	1.1.98.-
CCP43226.1	Long-chain-alcohol oxidase		1.1.3.20
CCP43640.1	NAD(P)/FAD-dependent oxidoreductase		1.14.13.-
CCP44538.1	L-gulonolactone oxidase	L-ascorbate biosynthesis	1.1.3.8
CCP45062.1	D-lactate dehydrogenase		1.1.99.6

4.3.2 Predictions for queries with new information regarding pathways:

Table 8. Orthologs found for the indicated query. The top panel identifies query code and previous annotations listing any identified pathway, Mycobrowser category and EC number. The bottom panel indicates ortholog's features retrieved from uniport and refer to hit description, up to two pathways and up to two EC numbers. The final column counts how many of these orthologs share the same characteristics. Tables 9-12 share this structure and refer to queries with lack of pathway information, that is, uncomplete or inexistent metabolic pathway inference or evidence that may be completed with RBH predictions.

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP42879.1	Acyl-CoA dehydrogenase FadE2		Lipid metabolism	1.3.8.7

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
Acyl-CoA dehydrogenase family member 11	Lipid metabolism; fatty acid beta-oxidation				88

CCP42879.1 from *M. tuberculosis* appears to be related to EC 1.3.8.7, a medium-chain acyl-CoA dehydrogenase and therefore a function in lipid metabolism can inferred for this protein. Fatty acid beta-oxidation is the pathway identified for all the 88 orthologs obtained for this protein. The alignments made have significant E-values from 7.01E-106 to 2.24E-149 therefore, have an accurate prediction for query's pathway.

Table 9

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP45916.1	NADPH-ferredoxin reductase FprA		Intermediary metabolism and respiration	1.18.1.2

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
ferredoxin-NADP(+) reductase		1.18.1.2			7
NADPH:adrenodoxin oxidoreductase, mitochondrial	Steroid metabolism; cholesterol metabolism	1.18.1.6			7
NADPH:adrenodoxin oxidoreductase, mitochondrial		1.18.1.6			28

Query's previous annotation's name coincides with the less numerous but present orthologs description ferredoxin-NADP(+) reductase annotations regarding FprA are experimental. The function suggested by the majority of orthologs is close to the already annotated and both proteins are part of the same pathway. Thus, the possible pathway in which this query is involved

is cholesterol metabolism and its function is predictably the one already identified (Fischer *et al.*, 2002). All alignments for the orthologs found are significant having E-values from 3.89E-76 to 9.09E-153 and identities of 36.1% to 52.4% being generally more significant for ferredoxin-NADP(+) reductase named orthologs.

Table 10

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP46093.1	Acyl-CoA dehydrogenase fadE25		Lipid metabolism	1.3.99.3

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
Acyl-CoA dehydrogenase, short-chain specific	Lipid metabolism; butanoate metabolism	1.3.8.1		1.3.8.1	2
Isovaleryl-CoA dehydrogenase, mitochondrial	Amino-acid degradation; L-leucine degradation; (S)-3-hydroxy-3-methylglutaryl-CoA from 3-isovaleryl-CoA: step 1/3	1.3.8.4			15
Short/branched chain specific acyl-CoA dehydrogenase, mitochondrial	Amino-acid degradation; L-isoleucine degradation	1.3.8.5	Lipid metabolism; mitochondrial fatty acid beta-oxidation		11
Medium-chain specific acyl-CoA dehydrogenase, mitochondrial	Lipid metabolism; mitochondrial fatty acid beta-oxidation	1.3.8.7			1
Glutaryl-CoA dehydrogenase	Aromatic compound metabolism; benzoyl-CoA degradation	1.3.99.32			1
Acyl-CoA dehydrogenase					1
Short-chain specific acyl-CoA dehydrogenase, mitochondrial					7

The EC number 1.3.99.3 for Acyl-CoA dehydrogenase fadE25 previously annotated for CCP46093.1, was transferred after our analysis to EC 1.3.8.7 for medium-chain acyl-CoA dehydrogenase, EC 1.3.8.8 for long-chain acyl-CoA dehydrogenase and EC 1.3.8.9 for very-long-chain acyl-CoA dehydrogenase. The only EC number related to the older one present on orthologs is 1.3.8.7, this number responds to medium chain Acyl-CoA. However, there are more possible catalytic functions for this protein since functions regarding short and branched chain specific dehydrogenases represented by 1.3.8.4, 1.3.8.5 and 1.3.8.1 EC numbers are more present in orthologs annotations, being 1.3.8.5 the most present and the most significant having E-values of 1E-100 in closely related species. All these named functions take part at least in two metabolic pathways: mitochondrial fatty acid beta-oxidation, part of the lipid metabolism and L-isoleucine and L-leucine degradation.

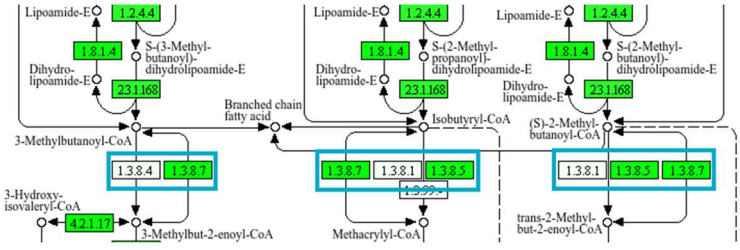

Figure 20. Valine, leucine and isoleucine degradation Pathway workflow with processes in terms of EC numbers, those present in Mycobacterium are marked in green. Image taken from KEGG PATHWAY Database

https://www.kegg.jp/pathway/mapno=00280&category=Mycobacterium. EC numbers from pred ictions made out of ortholog similarity are outlined in blue.

Like all prokaryotic microorganisms, *M. tuberculosis* bacteria do not have mitochondria, and theretofore mitochondrial fatty acid beta-oxidation cannot be contemplated as a possible pathway for this protein. However, some orthologs thought to possess this pathway belong to prokaryotic species and it is possible that this beta-oxidation occurs on the membrane. The broader pathway of lipid metabolism, as categorized in Mycobrowser, can be straightforward predicted from protein function. Isoleucine, leucine and valine degradation are also part of some ortholog's non-coincident catalytic activity. All functions present in orthologs are present in isoleucine, leucine and valine degradation pathways as shown in figure 20 so appear as plausible pathways where these enzymes can act.

Table 11. Summary table for queries with lack of pathway information inference or evidence that are completed with RBH predictions. Indicated information includes Query code, description, pathway and EC, further predictions are made in green. Predictions that indicated a more accurate function are taken in account for changing corresponding description.

Query code	Description	Pathways	EC
CCP42879.1	Acyl-CoA dehydrogenase FadE2	Lipid metabolism; fatty acid beta-oxidation	1.3.8.7
CCP46093.1	Acyl-CoA dehydrogenase fadE25 Short/branched chain specific acyl-CoA dehydrogenase	Lipid metabolism; fatty acid beta-oxidation / isoleucine, leucine and valine degradation	1.3.8.5
CCP45916.1	NADPH-ferredoxin reductase FprA	Cholesterol metabolism	1.18.1.2

4.3.3 Predictions for queries with new information regarding EC numbers and pathways:

Table 12. Orthologs found for the indicated query. The top panel identifies query code and previous annotations listing any identified pathway, Mycobrowser category and EC number. The bottom panel indicates ortholog's features retrieved from uniport and refer to hit description, up to two pathways and up to two EC numbers. The final column counts how many of these orthologs share the same characteristics. Tables 13-14 share this structure and refer to queries with lack of catalytic activity and pathway information, that is, uncomplete or inexistent EC number or metabolic pathway inference or evidence that may be completed with RBH predictions.

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP43601.1	α-keto-acid decaborxylase		Intermediary metabolism and respiration	4.1.1.-

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
pyruvate decarboxylase		4.1.1.1			34
branched-chain-2-oxoacid decarboxylase	Amino-acid degradation; Ehrlich pathway	4.1.1.72			1
Pyruvate decarboxylase					7
Similar to Saccharomyces cerevisiae YLR044C PDC1 Major of three pyruvate decarboxylase isozymes, key enzyme in alcoholic fermentation, decarboxylates pyruvate to acetaldehyde					2

Ortholog number is high, and the majoritarian ortholog protein name has a coincident annotation to that previously found, pyruvate metabolism pathway found is present in mycobacterium despite this protein lacks experimental evidence and is not predicted to be present (figure 21). E-values vary but in close species are found to be up to 9.43E-109.

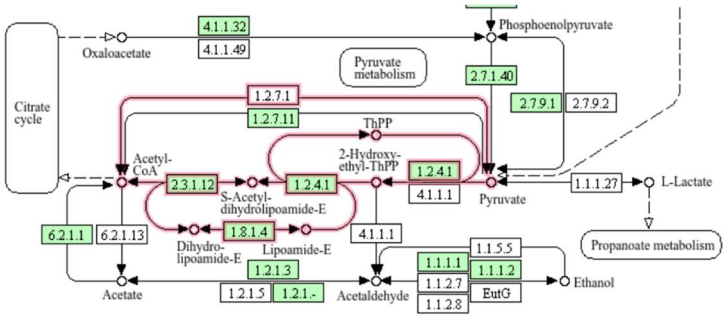

Figure 21. Pyruvate metabolism with steps inferred to be present in Mycobacterium tuberculosis are highlighted in green. Images are sourced from KEGG PATHWAY Database https://www.kegg.jp/pathway/mtu00010.

Table 13

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP45032.1	FAD-binding oxidoreductase		Intermediary metabolism and respiration	1.-.-.-

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
Alkylglycerone-phosphate synthase	Glycerolipid metabolism; ether lipid biosynthesis	2.5.1.26			1

Despite that there is only one ortholog for this protein, the function of Alkylglycerone-phosphate synthase is present in *M. tuberculosis* ether lipid metabolism (figure 22) and alignment responds to a significant 1.5E-49 E-value. Identity percentage is a 28% but the ortholog comes from a species of a eukaryote single-cell organism called *Stentor coeruleus* with a significative evolutionary distance.

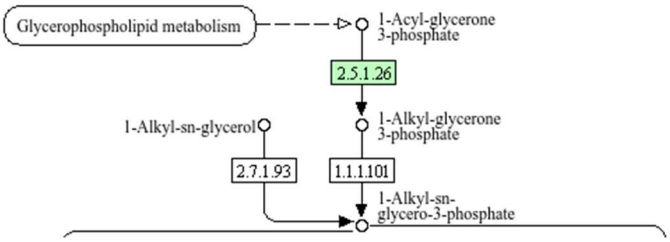

Figure 22. Ether metabolism pathway with steps inferred to be present for Mycobacterium tuberculosis marked in green, images taken from KEGG PATHWAY Database https://www.kegg.jp/pathway/mtu00565.

Table 14. Summary table for queries with lack of catalytic activity and pathway information inference or evidence that are completed with RBH predictions. Indicated information includes Query code, description, pathway and EC, further predictions are made in green. Predictions that indicated a more accurate function are taken in account for changing corresponding description.

Query code	Description	Pathway	EC
CCP43601.1	pyruvate decarboxylase	Carbohydrate metabolism; pyruvate metabolism	4.1.1.1
CCP44035.1	Oxigen dependent choline dehydrogenase	Amine and polyamine biosynthesis; betaine biosynthesis via choline pathway; betaine aldehyde from choline (cytochrome c reductase route): step 1/1	1.1.99.1
CCP45032.1	Alkylglycerone-phosphate synthase	Glycerolipid metabolism; ether lipid biosynthesis	2.5.1.26

4.4 Biological interpretation of UBH unknown protein results

Among the 117 queries with orthologs in UBH, 11 have orthologs providing new information only for EC numbers, 27 contain updates exclusively related to pathways, and 14 include both. These annotations may refine or override previous functional descriptions. However, as mentioned in section 4.3, many hits have low alignment significance (E-value > 1E-30), limiting their predictability. Tables 16–45 detail predictions and known annotations for cases with few, very functionally diverse, or low significance orthologs, while the remaining tables are provided in the supplementary material.

4.4.1 Predictions for queries with new information regarding EC numbers:

Table 15. Orthologs found for the indicated query. The top panel identifies query code and previous annotations listing any identified pathway, Mycobrowser category and EC number. The bottom panel indicates ortholog's features retrieved from uniport and refer to hit description, up to two pathways and up to two EC numbers. The final column counts how many of these orthologs share the same characteristics. Tables 16-23 share this structure and refer to queries with lack of catalytic activity information, that is, uncomplete or inexistent EC number inference or evidence that are completed with UBH predictions.

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP42887.1	FAD-binding oxidoreductase		Intermediary metabolism and respiration	1.-.-.-

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
D-2-hydroxyglutarate dehydrogenase		1.1.99.39			1

The only one ortholog has a function that coincides with the generally previous described for the query. The alignment presents a significant E-value score of 9.24E-67 and a 34.2% identity. Knowing that this homolog refers to *Xanthomonas citri pv. Viticola,* a very distant species, similarity is significant. Moreover, the ortholog's annotations are based on experimental evidence.

Table 16

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP43539.1	LLM class F420-dependent oxidoreductase		Conserved hypotheticals	1.-.-.- (1.14.14.5)*

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
Alkanesulfonate monooxygenase		1.14.14.5			4

Alignments are not significant for the orthologs of this query having E-values from 2.57E-09 to 7.2E-16. No further inference can be done with UBH annotations and EC number described as possible in previous annotations is most probably wrong.

Table 17

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP44118.1	LLM class F420-dependent oxidoreductase		Intermediary metabolism and respiration	1.14.-.-

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
5,10-methylenetetrahydromethanopterin reductase		1.5.98.2			1

Like for CCP42857.1 the only one ortholog Hit description refers to a function part of methanogenesis in which *M. tuberculosis* enzymes are not known to be implied. No further inferences can be done regarding to this query far from the ones already identified. An E-value of 5.04E-18 for the alignment is not significant enough and there is only one ortholog in the whole database.

Table 18

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP44704.1	FAD-binding oxidoreductase		Intermediary metabolism and respiration	1.-.-.-

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
Na(+)-translocating NADH-quinone reductase subunit F		7.2.1.1			25

Alignments are not significant for the orthologs of this query having E-values from 4.52E-09 to 3.97E-16. Also, length described for orthologs do not coincide with the one described for the query varying in almost 400 aminoacids.

Table 19

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP45863.1	NAD(P)H-dependent oxidoreductase		Conserved hypotheticals	1.-.-.-

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
NAD(P)H dehydrogenase (quinone)		1.6.5.2		1.6.5.2	1
2-hydroxy-1,4-benzoquinone reductase		1.6.5.7			1

Only two orthologs were found for this query but NAD(P)H dehydrogenase (quinone) from the bacteria *Burkholderia sp* has a lower E-value (3.18E-24 respect to a 1.57E-15) for the alignment and a 43% identity with almost 70% coverage, so its function is more probable. Previous general description for the query coincides with this predicted function.

Table 20

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP45888.1	LLM class F420-dependent oxidoreductase		Conserved hypotheticals	1.-.-.-

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
Alkanal monooxygenase		1.14.14.3			4
bacterial luciferase (Fragment)		1.14.14.3			12

Alignments are not significant for the orthologs of this query having E-values from 0.0000697 to 2.99E-07.

Table 21

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP46180.1	NADH:flavin oxidoreductase		Intermediary metabolism and respiration	1.5.-.-

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
NADPH dehydrogenase		1.6.99.1			2

Despite there is only two orthologs for this query, they are significant, having an E-value of 3.09E-27 and 4.56E-29 an identity of ~30% belonging to *Listeria* genus. Ortholog's function is similar to the one described by the annotations.

Table 22

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP46393.1	Flavin-dependent monooxygenase, oxygenase subunit HsaA	Steroid biosynthesis	Intermediary metabolism and respiration	1.14.14.12*

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
Dibenzothiophene monooxygenase	Sulfur metabolism; dibenzothiophene degradation	1.14.14.21			11

No further inferences can be done regarding to this query since E-values of the alignment are between 1.67E-09 and 9.35E-11.

Table 23. Summary table for queries with lack of catalytic activity information inference or evidence that are completed with UBH predictions. Indicated information includes Query code, description, pathway and EC, further predictions are made in green. Predictions that indicated a more accurate function are taken in account for changing corresponding description.

Query code	Description	Pathway	EC
CCP42887.1	D-2-hydroxyglutarate dehydrogenase		1.1.99.39
CCP44704.1	FAD-binding oxidoreductase		1.-.-.-
CCP46393.1	Flavin-dependent monooxygenase, oxygenase subunit HsaA	Steroid biosynthesis	1.14.14.12*
CCP43539.1	LLM class F420-dependent oxidoreductase		1.-.-.- ~~(1.14.14.5)*~~
CCP44118.1	LLM class F420-dependent oxidoreductase		1.14.-.-
CCP45888.1	LLM class F420-dependent oxidoreductase		1.-.-.-
CCP45863.1	NAD(P)H-dependent oxidoreductase		1.-.-.-
CCP46180.1	NADPH dehydrogenase		1.6.99.1

4.4.2 Predictions for queries with new information regarding pathways:

Table 24. Orthologs found for the indicated query. The top panel identifies query code and previous annotations listing any identified pathway, Mycobrowser category and EC number. The bottom panel indicates ortholog's features retrieved from uniport and refer to hit description, up to two pathways and up to two EC numbers. The final column counts how many of these orthologs share the same characteristics. Tables 25-32 share this structure and refer to queries with lack of catalytic activity information, that is, uncomplete or inexistent EC number inference or evidence that may be completed with UBH predictions.

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP42856.1	Acyl-CoA dehydrogenase FadE1		Lipid metabolism	1.3.8.7

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts

(R)-benzylsuccinyl-CoA dehydrogenase	Xenobiotic degradation; toluene degradation	1.3.8.3	1

Despite there's only one ortholog for the query has a significant alignment with an E-value 1.41E-54 with an alignment coverage of almost 90% and an identity of 34.8%, knowing that the species (*Thauera aromatica*) of the ortholog is distant from *M. tuberculosis*.

Table 25

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP42959.1	Acyl-CoA/acyl-ACP dehydrogenase FadE4		Lipid metabolism	1.3.8.7

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
glutaryl-CoA dehydrogenase (ETF)	Amino-acid metabolism; lysine degradation	1.3.8.6	Amino-acid metabolism; tryptophan metabolism		14

There is no significant alignments with this query they have E-values from 0.0000614 to 7.88E-07, despite orthologs are numerous.

Table 26

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP43115.1	Nitric oxide dioxygenase		Intermediary metabolism and respiration	1.14.12.17

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
nitric oxide dioxygenase		1.14.12.17		1.14.12.17	1
Putative methanesulfonate monooxygenase ferredoxin reductase subunit		1.14.13.111			1
Methane monooxygenase component C		1.14.13.25		1.14.13.25	1
Ferredoxin--NAD(P)(+) reductase CarAd		1.18.1.3		1.18.1.2	1
Naphthalene 1,2-dioxygenase/salicylate 5-hydroxylase systems, ferredoxin--NAD(P)(+), reductase component	Aromatic compound metabolism; naphthalene degradation	1.18.1.7		1.18.1.7	1
Na(+)-translocating NADH-quinone reductase subunit F		7.2.1.1			61
Aromatic O-demethylase, reductase subunit	Aromatic compound metabolism				1

Despite the most numerous ortholog's description is Na(+)-translocating NADH-quinone reductase subunit F, the lowest E-value obtained (6.01E-298) stands out significantly from the other alignments (2.02E-5 to 2.75E-24) and it is obtained for a inferred protein function of *Mycobacterium bovis* that coincides in description with the query previous annotations (nitric oxide dioxigenase). This significant alignment shares a 99.7% of identity. Also, this function is proven present in *M. tuberculosis* and *M. bovis* by experimental evidence in very similar genes for experimentally essayed gblN (Rv1542) (Ouellet et al.,) and our query (from gene Rv0385) (https://string-db.org/cgi/geneneighbors?taskId=b3snwz5s4Mra&sessionId=bCLJjKNdUBAA&node1=13980617&node2=13981807). No further annotations can be made for the query.

Table 27

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP43146.1	Glycine oxidase ThiO		Intermediary metabolism and respiration	1.4.3.19

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
D-amino-acid oxidase	Cofactor biosynthesis; thiamine diphosphate biosynthesis	1.4.3.3			2
tRNA 5-methylaminomethyl-2-thiouridine biosynthesis bifunctional protein MnmC		2.1.1.61			1

Orthologs have not significant alignments with the query with E-values of 3.15E-14, 3.63E-19 and 0.00000254. Hits' function do not coincide with the one previously described.

Table 28

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP44444.1	Acyl-CoA/acyl-ACP dehydrogenase FadE16		Lipid metabolism	1.3.99.3

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
Dibenzothiophene monooxygenase	Sulfur metabolism; dibenzothiophene degradation	1.14.14.21			2

One of the orthologs found for the query have a low E-value, while the other remain as not significant. Since orthologs are not experimentally annotated, no further inference about this query can be made. Previous annotations do not coincide with possible inference about the protein.

Table 29

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP44635.1	Ferredoxin reductase		Intermediary metabolism and respiration	1.18.1.2

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
Mercuric reductase		1.16.1.1			5
Chlorobenzene dioxygenase, ferredoxin reductase component	Aromatic compound metabolism	1.18.1.3			1
NADH-dependent phenylglyoxylate dehydrogenase subunit epsilon		1.2.1.58			1
NADH oxidase		1.6.3.4			7
monodehydroascorbate reductase (NADH)		1.6.5.4			10
Dihydrolipoyl dehydrogenase		1.8.1.4			8
lipoate--protein ligase	Protein modification; protein lipoylation via exogenous pathway; protein N(6)-(lipoyl)lysine from lipoate: step 1/2	6.3.1.20	Protein modification; protein lipoylation via exogenous pathway; protein N(6)-(lipoyl)lysine from lipoate: step 2/2		1
Apoptosis-inducing factor 1, mitochondrial					1

The most significant ortholog alignment is the one performed with chlorobenzene dioxygenase, ferredoxin reductase component with an E-value of 3.35E-54, 5 orders of magnitude inferior to the next most significant and coincident with previous annotations, leading to the possible pathway for the query and a secondary function. The ortholog's function is annotated from experimental evidence (Transformation of chlorinated benzenes and toulenes by Ralstonia sp., 2001)

Table 30

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP44701.1	Acyl-CoA dehydrogenase FadE17		Lipid metabolism	1.3.99.3

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
Isovaleryl-CoA dehydrogenase, mitochondrial (Fragment)	Amino-acid degradation; L-leucine degradation; (S)-3-hydroxy-3-methylglutaryl-CoA from 3-isovaleryl-CoA: step 1/3	1.3.8.4		1.3.8.1	1

The only one ortholog found doesn't have a significant alignment, with an E-value of 3.68E-23 and an identity percentage of 28.5% with a low query coverage.

Table 31

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP46626.1	Acyl-CoA dehydrogenase FadE35		Lipid metabolism	1.3.99.3

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
short-chain 2-methylacyl-CoA dehydrogenase	Amino-acid degradation; L-isoleucine degradation	1.3.8.5	Lipid metabolism		5
glutaryl-CoA dehydrogenase (ETF) (Fragment)	Amino-acid metabolism; lysine degradation	1.3.8.6	Amino-acid metabolism; tryptophan metabolism		1

Alignments for this protein orthologs are not significant enough having E-values of 2.68E-19 to 5.57E-21 with low (less than a 50%) query coverage.

Table 32. Summary table for queries with lack of pathway information inference or evidence that are completed with UBH predictions. Indicated information includes Query code, description, pathway and EC, further predictions are made in green. Predictions that indicated a more accurate function are taken in account for changing corresponding description.

Query code	Description	Pathway	EC
CCP43131.1	Acyl_CoA dehydrogenase FadE7 glutaryl-CoA dehydrogenase (ETF)	Amino-acid metabolism; lysine and tryptophan degradation	1.3.8.6
CCP42856.1	Acyl-CoA dehydrogenase FadE1 (R)-benzylsuccinyl-CoA dehydrogenase	Xenobiotic degradation; toluene degradation	1.3.8.3
CCP43621.1	Acyl-CoA dehydrogenase FadE10	Lipid metabolism; fatty acid beta-oxidation	1.3.8.7
CCP43721.1	Acyl-Coa dehydrogenase FadE12 Isovaleryl-CoA dehydrogenase	Amino-acid degradation; L-leucine degradation; (S)-3-hydroxy-3-methylglutaryl-CoA from 3-isovaleryl-CoA: step 1/3	1.3.8.4
CCP43724.1	Acyl-CoA dehydrogenase fadE13		1.3.8.7

CCP44701.1	Acyl-CoA dehydrogenase FadE17		1.3.99.3
CCP45294.1	Acyl-CoA dehydrogenase FadE19 (MMGC) short-chain 2-methylacyl-CoA dehydrogenase	Amino-acid degradation; L-isoleucine degradation/ Lipid metabolism; fatty acid beta-oxidation	1.3.8.5
CCP42879.1	Acyl-CoA dehydrogenase FadE2		1.3.8.7
CCP45522.1	Acyl-CoA dehydrogenase FadE20 Long-chain specific acyl-CoA dehydrogenase	Lipid metabolism; mitochondrial fatty acid beta-oxidation	1.3.8.8
CCP45951.1	Acyl-CoA dehydrogenase FadE23 short-chain 2-methylacyl-CoA dehydrogenase	Amino-acid degradation; L-isoleucine degradation/ Lipid metabolism; fatty acid beta-oxidation	~~1.3.8.7~~ 1.3.8.5
CCP45950.1	Acyl-CoA dehydrogenase FadE24		1.3.99.3
CCP46093.1	Acyl-CoA dehydrogenase fadE25		1.3.99.3
CCP42943.1	Acyl-CoA dehydrogenase FadE3 Isovaleryl-CoA dehydrogenase	Amino-acid degradation; L-leucine degradation; (S)-3-hydroxy-3-methylglutaryl-CoA from 3-isovaleryl-CoA: step 1/3	1.3.8.4
CCP46626.1	Acyl-CoA dehydrogenase FadE35		1.3.99.3
CCP43415.1	Acyl-CoA dehydrogenase FadE8		1.3.99.3
CCP43498.1	Acyl-CoA dehydrogenase FadE9 short-chain 2-methylacyl-CoA dehydrogenase	Amino-acid degradation; L-isoleucine degradation/ Lipid metabolism; fatty acid beta-oxidation	1.3.8.5
CCP44444.1	Acyl-CoA/acyl-ACP dehydrogenase FadE16		1.3.99.3
CCP45588.1	Acyl-CoA/acyl-ACP dehydrogenase FadE21 Isovaleryl-CoA dehydrogenase	Amino-acid degradation; L-leucine degradation; (S)-3-hydroxy-3-methylglutaryl-CoA from 3-isovaleryl-CoA: step 1/3	1.3.8.4
CCP42959.1	Acyl-CoA/acyl-ACP dehydrogenase FadE4		1.3.8.7
CCP44635.1	Ferredoxin reductase	Aromatic compound metabolism	1.18.1.2/1.18.1.3
CCP45029.1	Glycerol-3-phosphate dehydrogenase 1	Polyol metabolism; glycerol degradation via glycerol kinase pathway; glycerone phosphate from sn-glycerol 3-phosphate (aerobic route): step 1/1	1.1.5.3
CCP46121.1	Glycerol-3-phosphate dehydrogenase 2	Polyol metabolism; glycerol degradation	1.1.5.3
CCP43146.1	Glycine oxidase ThiO		1.4.3.19
CCP45916.1	NADPH-ferredoxin reductase FprA		1.18.1.2
CCP43115.1	Nitric oxide dioxygenase		1.14.12.17
CCP43944.1	Proline dehydrogenase	Amino-acid degradation; L-proline degradation into L-glutamate; L-glutamate from L-proline: step 1/2	1.5.5.2
CCP42977.1	Succinate dehydrogenase flavoprotein subunit [iron-sulfur subunit]	Carbohydrate metabolism; tricarboxylic acid cycle; fumarate from succinate (bacterial route): step 1/1	1.3.5.1 / 1.3.99.1

4.4.3 Predictions for queries with new information regarding EC numbers and pathways:

Table 33. Orthologs found for the indicated query. The top panel identifies query code and previous annotations listing any identified pathway, Mycobrowser category and EC number. The bottom panel indicates ortholog's features retrieved from uniport and refer to hit description, up to two pathways and up to two EC numbers. The final column counts how many of these orthologs share the same characteristics. Tables 39-46 share this structure and refer to queries with lack

of catalytic activity and pathway information, that is, uncomplete or inexistent EC number and metabolic pathway inference or evidence that may be completed with UBH predictions.

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP42785.1	FAD-binding oxidoreductase		Intermediary metabolism and respiration	1.-.-.-

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
Bifunctional solanapyrone synthase	Phytotoxin biosynthesis	1.1.3.42		5.5.1.20	1
Carbohydrate oxidase		1.1.3.5		1.1.3.5	1
FAD-linked oxidoreductase orf1	Mycotoxin biosynthesis				1

The most significant alignment was made with carbohydrate oxidase and has an E-value of 1.33E-27, coverage is high (90%) and has a 26.5% identity with the query. The function described by the ortholog might not be the same as the query but similar. Functions listed for this hit include various saccharide oxidations and are asserted by experimental evidence (Xu *et al.*, 2001).

Table 34

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP43180.1	NAD(P)/FAD-dependent oxidoreductase		Conserved hypotheticals	1.-.-.-

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
Protoporphyrinogen oxidase (Fragment)	Porphyrin-containing compound metabolism; protoporphyrin-IX biosynthesis; protoporphyrin-IX from protoporphyrinogen-IX: step 1/1	1.3.3.4			1

Only one hit was found with a non-significant alignment showing an E-value of 0.000263.

Table 35

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP43303.1	NAD(P)/FAD-dependent oxidoreductase		Intermediary metabolism and respiration	1.-.-.- (1.13.12.-)*

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
Flavin-containing monooxygenase	Plant hormone metabolism; auxin biosynthesis	1.14.13.168			1
FAD-dependent oxidoreductase (Fragment)		1.8.1.9			1
Flavin-containing monooxygenase					2

The few orthologs found do not give much information about the query, their alignments are not significant (0.000507 to 5.21E-12) despite are part of species from other kingdoms or bacterial phyla.

Table 36

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP43313.1	Oxidoreductase		Intermediary metabolism and respiration	1.-.-.-

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
3-amino-4-hydroxybenzoate 2-monooxygenase	Antibiotic biosynthesis	1.14.13.249			1
FAD-dependent monooxygenase cdmI	Secondary metabolite biosynthesis; terpenoid biosynthesis				1
Flavin-dependent monooxygenase					1

The homolog 3-amino-4-hydroxybenzoate 2-monooxygenase was the most significant alignment with an E-value of 3.43E-19 from *Streptomyces cremeus* bacterial species. This E-value is to high to make function inferences.

Table 37

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP43441.1	Mycofactocin system GMC family oxidoreductase		Intermediary metabolism and respiration	1.-.-.-

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
Paromamine 6'-oxidase	Antibiotic biosynthesis; neomycin biosynthesis	1.1.3.44		1.1.3.43	1
Cellobiose dehydrogenase		1.1.99.18			1

Both orthologs' alignments for paromamine 6'-oxidase and cellobiose dehydrogenase have 22.2, 24.2 identity percentage values and E-values of 1.1E-10, 9.07E-07 respectively belonging to *Phanerodontia chrysosporium* and *Streptomyces fradiae* bacteria, not presenting significant similarity.

Table 38

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP44016.1	Oxidoreductase		Intermediary metabolism and respiration	1.-.-.-

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
Zeaxanthin epoxidase, chloroplastic	Plant hormone biosynthesis; abscisate biosynthesis	1.14.15.21			9
Notoamide E oxidase notB	Alkaloid biosynthesis				1
Flavin-dependent monooxygenase					15

One of the orthologs named as flavin-dependent monooxygenases has an alignment with an E-value of 4.88E-79 of an unknown prokaryotic organism with evidence of activity involving 7-chlorotetracycline and tetracycline oxidation (Park *et al.*, 2017; Forsberg *et al.*, 2015). Other tetracycline monooxygenases have higher, not significant E-values for the alignments.

Table 39

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP44152.1	NAD(P)/FAD-dependent oxidoreductase		Intermediary metabolism and respiration	1.14.13.-

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
Senecionine N-oxygenase		1.14.13.101			1
Cyclopentanone 1,2-monooxygenase	Alcohol metabolism; cyclopentanol degradation; 5-valerolactone from cyclopentanol: step 2/2	1.14.13.16			1
Flavin-containing monooxygenase	Plant hormone metabolism; auxin biosynthesis	1.14.13.168			12
Pentalenolactone D synthase	Antibiotic biosynthesis; pentalenolactone biosynthesis	1.14.13.170			2
Acetone monooxygenase (methyl acetate-forming)		1.14.13.226			1
Thioredoxin reductase		1.8.1.9			2

Cyclopentanone 1,2-monooxygenase, Pentalenolactone D synthase, Pentalenolactone D synthase and Acetone monooxygenase (methyl acetate-forming) have E-values for the alignment with the query from 4.46E-51 to 1.69E-54. All of them are found in evolutionarily distant bacteria and its function is annotated by experimental evidence (Fordwour *et al.*, 2018; Reignier *et al.*, 2014). Cyclopentanone 1,2-monooxygenase is the most significant hit and therefore the nearest approach to the possible catalytic activity of the query. Note that the activity of the four significant orthologs consist of an oxidation of a ketone to obtain a lactone or an ester, a Baeyer–Villiger oxidation.

Table 40

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP44492.1	FAD-binding oxidoreductase		Intermediary metabolism and respiration	1.-.-.-

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
Reticuline oxidase	Alkaloid biosynthesis; (S)-scoulerine biosynthesis; (S)-scoulerine from (S)-reticuline: step 1/1	1.21.3.3			1
Tetrahydroberberine oxidase		1.3.3.8			2
(R)-6-hydroxynicotine oxidase	Alkaloid degradation; nicotine degradation; 6-hydroxypseudooxynicotine from nicotine (R-isomer route): step 2/2	1.5.3.6			1

Ortholog (R)-6-hydroxynicotine oxidase from *Paenarthrobacter nicotinovorans (Arthrobacter nicotinovorans)* species has the most significant alignment with an E-value of 5.31E-50. Its catalytic activity is asserted by evidence (Fitzpatrick *et al.*, 2016).

Table 41

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP44517.1	FAD-dependent oxidoreductase		Intermediary metabolism and respiration	1.-.-.-

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts

Aklavinone 12-hydroxylase DnrF	Antibiotic biosynthesis; daunorubicin biosynthesis	1.14.13.180	Antibiotic biosynthesis; carminomycin biosynthesis	1
6-methylpretetramide 4-monooxygenase	Antibiotic biosynthesis; oxytetracycline biosynthesis	1.14.13.232		1
Kynurenine 3-monooxygenase	Cofactor biosynthesis; NAD(+) biosynthesis; quinolinate from L-kynurenine: step 1/3	1.14.13.9		1
Squalene monooxygenase	Terpene metabolism; lanosterol biosynthesis; lanosterol from farnesyl diphosphate: step 2/3	1.14.14.17		24
Dialkyldecalin synthase	Antibiotic biosynthesis			1

Alignment for 6-methylpretetramide 4-monooxygenase 1.14.13.232 in *Streptomyces rimosus* has a significant E-value of 3.03E-50, its protein function is asserted by experimental evidence (Wang *et al.*, 2009).

Table 42

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP45575.1	PDR/VanB family oxidoreductase		Intermediary metabolism and respiration	1.-.-.-

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
nitric oxide dioxygenase		1.14.12.17		1.14.12.17	4
Carnitine monooxygenase reductase subunit	Amine and polyamine metabolism; carnitine metabolism	1.14.13.239		1.14.13.239	1
NADH-cytochrome b5 reductase		1.6.2.2			1

The most significant ortholog was carnitine monooxygenase reductase subunit from *Acinetobacter sp. 39-4* species, with an E-value of 1.16E-41 with a 100% alignment coverage and a 28.1% identity.

Table 43

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP45858.1	NAD(P)/FAD-dependent monooxygenase		Intermediary metabolism and respiration	1.-.-.- (1.13.12.-)*

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
Flavin-containing monooxygenase	Plant hormone metabolism; auxin biosynthesis	1.14.13.168			3
L-ornithine N(5)-monooxygenase	Siderophore biosynthesis	1.14.13.196		1.14.13.196	2

Orthologs with L-ornithine N(5)-monooxygenase description present the most significant alignments with E-values of 2.09E-84 and 3.03E-85 for *Periglandula ipomoeae* and *Epichloe inebrians* (fungi) species, being coherent with general previous annotations for the query.

Table 44

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP46342.1	LLM class F420-dependent oxidoreductase		Intermediary metabolism and respiration	1.14.-.-

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
5,10-methylenetetrahydromethanopterin reductase	One-carbon metabolism; methanogenesis from CO(2); methyl-coenzyme M from 5,10-methylene-5,6,7,8-tetrahydromethanopterin: step 1/2	1.5.98.2			1

The only one ortholog found is implied in methanogenesis and has a low significance E-value of 5.25E-20.

Table 45

Query code	Description	Pathway	MYCOBROWSER Category	EC
CCP46658.1	NAD(P)/FAD-dependent oxidoreductase		Intermediary metabolism and respiration	1.-.-.-

Hit description	Pathway 1	EC 1	Pathway 2	EC 2	counts
Phytoene desaturase (lycopene-forming)	Carotenoid biosynthesis; astaxanthin biosynthesis	1.3.99.31			2
prolycopene isomerase	Carotenoid biosynthesis; lycopene biosynthesis	5.2.1.13			23

With E-values of 0.000528 to 3.14E-11 no significant alignments are found for this query despite there are numerous orthologs.

Table 46. Summary table for queries with lack of pathway and catalytic activity information inference or evidence that are completed with UBH predictions. Indicated information includes Query code, description, pathway and EC, further predictions are made in green. Predictions that indicated a more accurate function are taken in account for changing corresponding description.

Query code	Description	Pathway	EC
CCP42785.1	Carbohydrate oxidase	Energy metabolism	(1.1.3.5)*
CCP43180.1	NAD(P)/FAD-dependent oxidoreductase		1.-.-.-
CCP43226.1	Long-chain-alcohol oxidase	Energy metabolism	1.1.3.20/1.1.3.13
CCP43303.1	NAD(P)/FAD-dependent oxidoreductase		1.-.-.- (1.13.12.-)*
CCP43313.1	Oxidoreductase		1.-.-.-
CCP43441.1	Mycofactocin system GMC family oxidoreductase		1.-.-.-
CCP44016.1	flavin-dependent monooxygenase (7-chlorotetracycline and tetracycline oxidation)		1.-.-.-
CCP44152.1	Cyclopentanone 1,2-monooxygenase	Alcohol metabolism; cyclopentanol degradation; 5-valerolactone from cyclopentanol: step 2/2	1.14.13.16
CCP44492.1	FAD-binding oxidoreductase (R)-6-hydroxynicotine oxidase	Alkaloid degradation; nicotine degradation; 6-hydroxypseudooxynicotine from nicotine (R-isomer route): step 2/2	1.5.3.6
CCP44517.1	6-methylpretetramide 4-monooxygenase	Antibiotic biosynthesis; oxytetracycline biosynthesis	1.14.13.232
CCP45575.1	Carnitine monooxygenase reductase subunit	Amine and polyamine metabolism; carnitine metabolism	1.14.13.239

CCP45858.1	L-ornithine N(5)-monooxygenase	Siderophore biosynthesis	1.14.13.196
CCP46342.1	LLM class F420-dependent oxidoreductase		1.14.-.-
CCP46658.1	NAD(P)/FAD-dependent oxidoreductase		1.-.-.-

4.5 Metabolic context and pharmaceutical and biotechnological applications of newly inferred protein functions.

Among all queried and reviewed (54) proteins 33 are found to retrieve new reliable functional information. Long-chain-alcohol oxidase (CCP43226.1), (R)-6-hydroxynicotine oxidase (CCP44492.1) and NADPH dehydrogenase (CCP46180.1) are known to be present in 29 species of Mycobacterium and not present in any of the 8 mammal species revised including homo sapiens (Montesa et al., 2023 TFG). These two characteristics reveal drug target potential yet enzymes are involved in core proteins of all genus species, believed to play a crucial role on bacteria survival and not present in humans. All new functionally characterized proteins bind to FAD cofactor except carnitine monooxygenase and NADPH dehydrogenase that bind FMN. New functionally described proteins are involved in amino-acid metabolism, lipid metabolism, xenobiotic and aromatic compound degradation, antibiotic biosynthesis biosynthesis, energy metabolism and amine synthesis. Further descriptions and other biotechnological and pharmaceutical applications are described in more detail for each metabolic route.

Proteins involving amino-acid degradation and metabolism:

- Glutaryl-CoA dehydrogenase (CCP43131.1) that catalyzes the oxidative decarboxylation of glutaryl-CoA to (2E)-butenoyl-CoA encoded by FadE7 gene variant, is involved in lysine and tryptophan degradation.

- Isovaleryl-CoA dehydrogenase encoded by genes FadE12, FadE3 and FadE21 (CCP43721.1, CCP42943.1, CCP45588.1) implied in leucine degradation and catalyzer of the oxidation from isovaleryl-CoA (3-methylbutanoyl-CoA) to 3-methyl-(2E)-butenoyl-CoA.

- Proline dehydrogenase (CCP43944.1) degrades and oxidizes proline to glutamate.

- L-gulonolactone oxidase is implied in ascorbate synthesis from glucose where L-gulono-1,4-lactone is oxidized to L-ascorbate. This is a protein newly involving amino-acid synthesis.

Proteins implied in fatty acid beta-oxidation and lipid metabolism:

- Short-chain 2-methylacyl-CoA dehydrogenase encoded by gene variants FadE19, FadE23, FadE9 (CCP45294.1, CCP45951.1, CCP43498.1), it catalyzes 2-methylbutanoyl-CoA oxidation to (2E)-2-methylbut-2-enoyl-CoA in fatty acid beta-oxidation.

- Short/branched chain specific acyl-CoA (CCP46093.1) encoded by gene variant FadE25 is involved in fatty acid beta-oxidation in which preferably (2R)-2-methylbutanoyl-CoA is oxidized to ethylacryloyl-CoA. Possible cholesterol side chain degrader that would lead to a potential drug target, an acute immune response of TB host is lipid metabolism gene up-regulation, mycobacteria use steroids as a primary source for energy and thus, survival (Wilburn *et al.*, 2018; Wipperman *et al.*, 2014).

- Long-chain acyl-CoA dehydrogenase (CCP45522.1) encoded by gene variant FadE20 oxidizes preferably (5Z)-tetradecenoyl-CoA to (2E,5Z)-tetradecadienoyl-CoA participating in fatty acid beta-oxidation.

- Alkylglycerone-phosphate synthase (CCP45032.1) participates in ether lipid biosynthesis oxidizing a log chain fatty alcohol into a 1-O-alkylglycerone 3-phosphate.

- Long-chain-alcohol oxidase (CCP43226.1) involved in lipid oxidation, oxidizing a long-chain primary fatty alcohol into its corresponding aldehyde.

Proteins participating in xenobiotic and aromatic compound degradation:

- (R)-benzylsuccinyl-CoA dehydrogenase is aprotein encoded by gene variant FadE1 (CCP42856.1) participates in toluene degradation by oxidizing (R)-benzylsuccinyl-CoA to (E)-2-benzylidenesuccinyl-CoA. Useful for toluene degradation since anaerobic oxidation of petroleum hydrocarbons can be coupled to the reduction of metals, this will accelerate the removal of pollutants(Tremblay and Zhang, 2017).

- D-2-hydroxyglutarate dehydrogenase (CCP42887.1) EC number suggest that might be involved in methylglyoxal pathway degrading and oxidizing the subproduct of glycolysis (R)-2-hydroxyglutarate to 2-oxoglutarate (Engqvist *et al.*, 2009).

- Cyclopentanone 1,2 monooxygenase (CCP44152.1) implied cyclopentanol degradation by oxidizing cyclopentanone into 5-valerolactone. It is useful for biocatalysis in Baeyer-Villiger reactions for the synthesis of bioactive products (Reignier *et al.*, 2014).

- Flavin-dependent monooxygenase (7-chlorotetracycline monooxygenase) (CCP44016.1) hydroxylates 7-chlorotetracycline and/or tetracycline and participates in antibiotic degradation.

- (R)-6-hydroxynicotine oxidase (CCP44492.1) participates in alkaloid degradation by nicotine degradation, obtaining 6-hydroxypseudooxynicotine from nicotine by R-isomer route.

- Ferredoxin reductase (CCP44635.1) is an electron transfer in aromatic metabolism for [2Fe-2S]-[ferredoxin] oxidation having a chlorobenzene dioxygenase subunit.

Antibiotic biosynthesis implicated proteins:

- 6-methylpretetramide 4-monooxygenase (CCP44517.1) oxidates 6-methylpretetramide to 4-hydroxy-6-methylpretetramide participating in oxytetracycline biosynthesis (antibiotic biosynthesis).

- NAD(P)/FAD-dependent monooxygenase (L-ornithine N(5)) (CCP45858.1) participates in siderophore biosynthesis by oxidating L-ornithine to N(5)-hydroxy-L-ornithine.

- Carnitine monooxygenase reductase subunit (CCP45575.1) reduces carnitine to (3R)-3-hydroxy-4-oxobutanoate with trimethylamine as a subproduct participating in amine and polyamine metabolism.

Energy metabolism and amine synthesis:

- Carbohydrate oxidase (CCP42785.1) participates in general energy metabolism by saccharide oxidation.

- Glycerol-3-phosphate dehydrogenase 1 & 2 (CCP45029.1, CCP46121.1) participates in polyol metabolism by glycerol degradation (energy metabolism and other purposes). The protein oxidizes glycerol 3-phosphate to dihydroxyacetone phosphate.

- NADPH dehydrogenase (CCP46180.1) serves as an electron transfer in energy metabolism.

- Oxygen-dependent choline dehydrogenase (CCP44035.1) is involved in betaine pathway for amine and polyamine biosynthesis. The protein obtains betaine aldehyde from choline.

- Pyruvate decarboxylase (CCP43601.1) decarboxylates pyruvate to acetaldehyde, useful for ethanol synthesis (Tian et al., 2017).

- Succinate dehydrogenase flavoprotein subunit (CCP42977.1) participates in the tricarboxylic acid cycle as part of carbohydrate metabolism converting succinate to fumarate.

5 Conclusions

- A novel computational method has been developed to effectively search for new catalytic activities in flavoenzymes within the *Mycobacterium tuberculosis* flavoproteome.

- The method reliably predicted 33 flavoprotein new complete functions leading to more accurate and metabolically contextualized descriptions, spanning diverse pathways and leading to 2 potential applications in biocatalysis and 3 in drug targeting.

- It achieved approximately 60% agreement with previous annotations and successfully analyzed 184 proteins in around 50 minutes.

- Of the 133 unknown proteins, 54 were found to have similarity with fully annotated flavoproteins from all available species comprehending all kingdoms. 33 queries were found to have significance in HSPs alignments.

Bibliography

World Health Organization (2024) Global Tuberculosis Report 2024. Geneva: World Health Organization; Available from: https://www.who.int/teams/global-programme-on-tuberculosis-and-lung-health/tb-reports/global-tuberculosis-report-2024

Abbasi Mesrabadi,H. et al. (2023) Drug–target interaction prediction based on protein features, using wrapper feature selection. Sci Rep, 13:3594.

Al-Dalky,R. et al. (2016) Applying Monte Carlo Simulation to Biomedical Literature to Approximate Genetic Network. IEEE/ACM Trans Comput Biol Bioinform, 13, 494–504.

Aleksander,S.A. et al. (2023) The Gene Ontology knowledgebase in 2023. Genetics, 224.

Altenhoff,A.M. et al. (2012) Resolving the ortholog conjecture: Orthologs tend to be weakly, but significantly, more similar in function than paralogs. PLoS Comput Biol, 8:1002514.

Altenhoff,A.M. et al. (2016) Standardized benchmarking in the quest for orthologs. Nat Methods, 13, 425–430.

Altschup,S.F. et al. (1990) Basic Local Alignment Search Tool, 215, 403–410.

Andreeva,A. et al. (2020) The SCOP database in 2020: Expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Res, 48, D376–D382.

Ashkenazy,H. et al. (2019) Multiple Sequence Alignment Averaging Improves Phylogeny Reconstruction. Syst Biol, 68, 117–130.

Bark,C.M. et al. (2024) 14:11 Annual Review of Medicine Downloaded from www.annualreviews.org. Guest (guest) IP: 84.232.25.73 On: Fri. 14, 11.

Bartont,G.J. and Sternberg,M.J. A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparisons. J Mol Biol. 198, 327-37.

Bateman,A. et al. (2023) UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res, 51, D523–D531.

Buchfink,B. et al. (2014) Fast and sensitive protein alignment using DIAMOND. Nat Methods, 12, 59–60.

Buchfink,B. et al. (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods, 18, 366–368.

Chao,J. et al. (2022) Developments in Algorithms for Sequence Alignment: A Review. Biomolecules, 12:546.

Cremades,N. et al. (2009) Discovery of specific flavodoxin inhibitors as potential therapeutic agents against Helicobacter pylori infection. ACS Chem Biol, 4, 928–938.

Dalquen,D.A. and Dessimoz,C. (2013) Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals. Genome Biol Evol, 5, 1800–1806.

Damkliang,K. et al. Similarity Score Estimation and Gaps Trimming of Multiple Sequence Alignment for Phylogenetic Tree Analysis, 11, 129-142.

Montesa, A. et al. Mycobacterium tuberculosis (2023) TFM

Eggers,R. et al. (2021) The scope of flavin-dependent reactions and processes in the model plant Arabidopsis thaliana. Phytochemistry, 189:112822.

Engqvist,M. et al. (2009) Two D-2-hydroxy-acid dehydrogenases in arabidopsis thaliana with catalytic capacities to participate in the last reactions of the methylglyoxal and β-oxidation pathways. Journal of Biological Chemistry, 284, 25026–25037.

Farhat,M. et al. (2024) Drug-resistant tuberculosis: a persistent global health concern. Nat Rev Microbiol, 22, 617–635.

Fischer,F. et al. (2002) Mycobacterium tuberculosis FprA, a novel bacterial NADPH-ferredoxin reductase. Eur J Biochem, 269, 3005–3013.

Fitzpatrick,P.F. et al. (2016) Mechanism of the Flavoprotein l -Hydroxynicotine Oxidase: Kinetic Mechanism, Substrate Specificity, Reaction Product, and Roles of Active-Site Residues. Biochemistry, 55, 697–703.

Fordwour,O.B. et al. (2018) Kinetic characterization of acetone monooxygenase from Gordonia sp. strain TY-5. AMB Express, 8, 181.

Forsberg,K.J. et al. (2015) The Tetracycline Destructases: A Novel Family of Tetracycline-Inactivating Enzymes. Chem Biol, 22, 888–897.

Di Franco,A. et al. (2019) Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. BMC Evol Biol, 19:21.

Gonzalez,M.W. and Pearson,W.R. (2010) Homologous over-extension: A challenge for iterative similarity searches. Nucleic Acids Res, 38, 2177–2189.

Graef,J. et al. (2022) Searching Geometric Patterns in Protein Binding Sites and Their Application to Data Mining in Protein Kinase Structures. J Med Chem, 65, 1384–1395.

Gudipati,V. et al. (2014) The flavoproteome of the yeast Saccharomyces cerevisiae. Biochim Biophys Acta Proteins Proteom, 1844, 535–544.

Gusfield,Dan. (1997) Algorithms on strings, trees, and sequences : computer science and computational biology Cambridge University Press. p. i–vi.

Hu,B. et al. (2018) Feature Selection for Optimized High-Dimensional Biomedical Data Using an Improved Shuffled Frog Leaping Algorithm. IEEE/ACM Trans Comput Biol Bioinform, 15, 1765–1773.

Hu,Q.N. et al. (2012) Assignment of EC Numbers to Enzymatic Reactions with Reaction Difference Fingerprints. PLoS One, 7(12):e52901

Iwata,H. et al. (2013) Inferring protein domains associated with drug side effects based on drug-target interaction network. BMC Syst Biol, 7 Suppl 6.

Jackson,M. et al. (2018) The genetic basis of disease. Essays Biochem, 62, 643–723.

Jaeger,S. et al. (2008) Integrating protein-protein interactions and text mining for protein function prediction. In, BMC Bioinformatics 9:S2.

Jumper,J. et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature, 596, 583–589.

Kerfeld,C.A. and Scott,K.M. (2011) Using BLAST to teach 'E-value-tionary' concepts. PLoS Biol, 9:e1001014.

Koonin,E. V. (2005) Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet, 39, 309–338.

Lienhart,W.D. et al. (2013) The human flavoproteome. Arch Biochem Biophys, 535, 150–162.

Lin,B. et al. (2024) A comprehensive review and comparison of existing computational methods for protein function prediction. Brief Bioinform, 25, Issue 4, bbae289.

MacHeroux,P. et al. (2011) Flavogenomics - A genomic and structural view of flavin-dependent proteins. FEBS Journal, 278, 2625–2634.

Mansour,A. et al. Assessment of Molecular (Dis)similarity: The Role of Multiple Sequence Alignment (MSA) Programs in Biological Research 3, 23–30.

Massey,V. (1995) Introduction: Flavoprotein structure and mechanism. The FASEB Journal, 9, 473–475.

Massey',V. (2000) The Chemical and Biological Versatility of Riboflavin, 28, 283-96.

McGinnis,S. and Madden,T.L. (2004) BLAST: At the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res, 32, W20-5.

Minjárez-Sáenz,M. et al. (2022) Mining the Flavoproteome of Brucella ovis, the Brucellosis Causing Agent in Ovis aries. Microbiol Spectr, 10:e0229421.

Navarro,Gonzalo. and Raffinot,Mathieu. (2002) Flexible pattern matching in strings : practical online search algorithms for texts and biological sequences Cambridge University Press.

Newman,A.M. and Cooper,J.B. (2010) AutoSOME: A clustering method for identifying gene expression modules without prior knowledge of cluster number. BMC Bioinformatics, 11:117.

Nguyen,B. et al. (2020) Preventive Therapy for Multidrug Resistant Latent Tuberculosis Infection: An Ethical Imperative with Ethical Barriers to Implementation, 19–35.

Otto-Knapp,R. et al. (2024) Availability of drugs for the treatment of multidrug-resistant/rifampicin-resistant tuberculosis in the World Health Organization European Region, October 2023. Eurosurveillance, 29:2400211.

Ouellet,H. et al. Truncated hemoglobin HbN protects Mycobacterium bovis from nitric oxide, 99, 5902-5907.

Park,J. et al. (2017) Plasticity, dynamics, and inhibition of emerging tetracycline resistance enzymes. Nat Chem Biol, 13, 730–736.

Pearson,W.R. (2013) An introduction to sequence similarity ('homology') searching. Curr Protoc Bioinformatics 3:3.1.1-3.1.8.

Reignier,T. et al. (2014) Broadening the scope of Baeyer–Villiger monooxygenase activities toward α,β-unsaturated ketones: A promising route to chiral enol-lactones and ene-lactones. Chemical Communications, 50, 7793–7796.

Rotondo,A. and Quilligan,F. (2020) Evolution Paths for Knowledge Discovery and Data Mining Process Models. SN Comput Sci, 1:109.

Schlicker,A. et al. (2006) A new measure for functional similarity of gene products based on gene ontology. BMC Bioinformatics, 7:302.

Sindhu,S. and Sindhu,D. (2017) International Journal of Computer Science and Mobile Computing Data Mining and Gene Expression Analysis in Bioinformatics, 6, 72-83.

Singh,P. and Singh,N. (2021) Role of Data Mining Techniques in Bioinformatics. International Journal of Applied Research in Bioinformatics, 11, 51–60.

Siu,W.-Y. et al. (2010) A data-mining approach for multiple structural alignment of proteins. print) Bioinformation, 4, 366–370.

Sivashankari,S. and Shanmughavel,P. (2006) Hypothesis Functional annotation of hypothetical proteins-A review, 1, 335–338.

Stamboulian,M. et al. (2020) The ortholog conjecture revisited: The value of orthologs and paralogs in function prediction. Bioinformatics, 36, I219–I226.

Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol, 35, 1026–1028.

Stringer,B. et al. (2023) Function Prediction.

Surya Narayana,G. and Vasumathi,D. (2018) An Attributes Similarity-Based K-Medoids Clustering Technique in Data Mining. Arab J Sci Eng, 43, 3979–3992.

Thareja,P. and Chhillar,R.S. (2020) A review of data mining optimization techniques for bioinformatics applications. International Journal of Engineering Trends and Technology, 68, 58–62.

Tian,L. et al. (2017) Enhanced ethanol formation by Clostridium thermocellum via pyruvate decarboxylase. Microb Cell Fact, 16, 171.

Tremblay,P.-L. and Zhang,T. (2017) Functional Genomics of Metal-Reducing Microbes Degrading Hydrocarbons. In, Anaerobic Utilization of Hydrocarbons, Oils, and Lipids. Springer International Publishing, 1–21.

Miura,R. (2001) Versatility and Specificity in Flavoenzymes: Control Mechanisms of Flavin Reactivity. The Chemical Record , 1, 183-194.

Wang,P. et al. (2009) Identification of oxyE as an ancillary oxygenase during tetracycline biosynthesis. ChemBioChem, 10, 1544–1550.

Wilburn,K.M. et al. (2018) Cholesterol and fatty acids grease the wheels of Mycobacterium tuberculosis pathogenesis. Pathog Dis, 76.

Wipperman,M.F. et al. (2014) Pathogen roid rage: Cholesterol utilization by Mycobacterium tuberculosis. Crit Rev Biochem Mol Biol, 49, 269–293.

Wolf,Y.I. and Koonin,E. V. (2012) A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. Genome Biol Evol, 4, 1286–1294.

Xu,F. et al. (2001) A novel carbohydrate:acceptor oxidoreductase from Microdochium nivale. Eur J Biochem, 268, 1136–1142.

Xu,X. and Bonvin,A.M.J.J. (2024) DeepRank-GNN-esm: A graph neural network for scoring protein-protein models using protein language model. Bioinformatics Advances, 4, Issue 1, vbad191.

Zhang,C. et al. (2020) Riboflavin Is Directly Involved in N-Dealkylation Catalyzed by Bacterial Cytochrome P450 Monooxygenases. ChemBioChem, 21, 2297–2305.

Zhang,Z. et al. (1998) Protein sequence similarity searches using patterns as seeds 26, 3986-90.

Zhong,F. et al. Drug target inference by mining transcriptional data using a novel graph convolutional network framework, 13, 281-301