OXFORD
UNIVERSITY PRESS | DATABASE

Mabellini: a genome-wide database for understanding the structural proteome and evaluating prospective antimicrobial targets of the emerging pathogen *Mycobacterium abscessus*

Journal:	*DATABASE*
Manuscript ID	DATABASE-2019-0053.R2
Manuscript Type:	Original Article
Date Submitted by the Author:	27-Aug-2019
Complete List of Authors:	Skwark, Marcin; University of Cambridge, Department of Biochemistry Torres, Pedro Henrique; University of Cambridge, Department of Biochemistry Copoiu, Liviu; University of Cambridge, Department of Biochemistry Bannerman, Bridget; University of Cambridge, Department of Biochemistry Floto, R. Andres; University of Cambridge, Department of Medicine; Royal Papworth Hospital, Cambridge Centre for Lung Infection Blundell, Tom L. ; University of Cambridge
Keywords:	Mycobacterium abscessus, Database, Structural Proteome, Protein Modelling, Mycobacteria
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
HieraGo.py ProteGo.py	

Mabellini: a genome-wide database for understanding the structural proteome and evaluating prospective antimicrobial targets of the emerging pathogen *Mycobacterium abscessus*

Marcin J. Skwark[1¶], Pedro H. M. Torres[1¶], Liviu Copoiu[1¶], Bridget Bannerman[1], R. Andres Floto[2,3] and Tom L. Blundell[1*]

[1]*Department of Biochemistry, University of Cambridge, Cambridge, CB2 1GA, UK*

[2]*Molecular Immunity Unit, Department of Medicine University of Cambridge, MRC-Laboratory of Molecular Biology, Cambridge, CB2 0QH, UK*

[3]*Cambridge Centre for Lung Infection, Royal Papworth Hospital, Cambridge, CB23 3RE, UK*

*Corresponding Author
E-mail: tlb20@cam.ac.uk (TLB)

¶These authors have contributed equally to this manuscript

Abstract

Mycobacterium abscessus, a rapid growing, multidrug resistant, nontuberculous mycobacteria (NTM), can cause a wide range of opportunistic infections, particularly in immunocompromised individuals. *M. abscessus* has emerged as a growing threat to patients with cystic fibrosis (CF), where it causes accelerated inflammatory lung damage, is difficult and sometimes impossible to treat, and can prevent safe transplantation. There is therefore an urgent unmet need to develop new therapeutic strategies. The elucidation of the *M. abscessus* genome in 2009 opened a wide range of research possibilities in the field of drug discovery that can be more effectively exploited upon the characterization of the structural proteome. Where there are no experimental structures, we have used the available amino acid sequences to create 3D models of the majority of the remaining proteins that constitute the *M. abscessus* proteome (3,394 proteins and over 13,000 models) using a range of up-to-date computational tools, many developed by our own group. The models are freely available for download in an on-line database, together with quality data and functional annotation. Furthermore, we have developed an intuitive and user-friendly web interface (http://www.mabellinidb.science) that enables easy browsing, querying and retrieval of the proteins of interest. We believe that this resource will be of use in evaluating the prospective targets for design of antimicrobial agents and will serve as a cornerstone to support the development of new molecules to treat *M. abscessus* infections.

Database URL: http://www.mabellinidb.science

Introduction

Mycobacterium abscessus is a species of rapid growing mycobacteria (RGM), comprised of three subspecies(1,2) (*abscessus*, *massiliense*(3) and *boletti*(4)) collectively referred to as *M. abscessus Complex*(1).

While most RGM, such as *M. smegmatis* and *M. marinum,* are saprophytic organisms, either found as free-living organisms or intracellular parasites, mainly infecting amoebae(5), data regarding environmental isolates of *M. abscessus* are scarce(6). In fact, it is one of the few RGM capable of infecting human hosts, causing skin and pulmonary diseases, and is frequently (but not exclusively) isolated from patients with pre-existing underlying diseases, such as cystic fibrosis (CF)(6). For CF patients, *M. abscessus* infections are particularly troublesome, causing an accelerated inflammatory lung damage, evidenced by rapid decline in lung function(7) and frequently preventing safe lung transplantation(8,9). Unlike other nontuberculous mycobacteria (NTM), *M. abscessus* can be transmitted from patient to patient(10), probably through exposure to infectious aerosols and/or fomites(11).

Treating *M. abscessus* infections is extremely challenging due to (i) intrinsic resistance to antibiotics, such as isoniazid, rifampin, pyrazinamide and ethambutol, used to treat *M. tuberculosis*(6); (ii) acquired drug resistance, such as resistance to aminoglycosides via mutations in the 16S ribosomal RNA(12,13); (iii) inducible tolerance, such as in the case of clarithromycin, a macrolide drug, where resistance is likely to occur through the methylation of ribosomal RNA carried out by the erythromycin ribosomal methylase Erm(41)(14,15); and (iv) survival within antibiotic-inaccessible niches (such as within biofilms)(16). As a consequence, treatment failure is extremely high (60-70% in individuals with CF(17)), with patients experiencing persistent culture positivity, relapse following initial sputum culture conversion(18–20), or early recurrence after the end of treatment(21).

Recently, Luthra and co-workers(22) have reviewed the genotypic determinants of antibiotic resistance in *M. abscessus* and pinpointed several proteins, which act through distinct mechanisms and are likely responsible for the emergence of resistant phenotypes; we revisit some of this in the results section. Besides antibiotic resistance, other phenotypic aspects related to virulence have been intensively studied in the past decades, but many underlying mechanisms involved in *M. abscessus* pathogenesis are still not completely understood.

M. abscessus can grow with two distinct colony morphologies, namely smooth (S) and rough (R) phenotypes(23,24). M. abscessus isolates with a R morphotype have been associated with increased pathogenicity(25). It was recently shown that the conversion from S to R phenotype can occur during the course of an infection and is caused by loss of Glycopeptidolipid (GPL)(24), due to mutations taking place in the GPL locus(26). This is an interesting case in which loss of function leads to selective advantage and, ultimately, to increased pathogenicity.

Another interesting hallmark associated with R isolates(27) is their ability to form cord-like structures, in which the long axes of the bacilli align to the horizontal axis of the cord, a trait shared by several other mycobacterial species.(28) It is believed that this kind of organization is directly linked to the pathogenicity, hindering effective phagocytic clearance(29), enhancing the persistence of bacterial cells inside macrophages(28) and causing extensive macrophage damage(30), Llorens-Fons and co-workers have presented evidence that Trehalose Polyphleates (TPP) are crucial molecules for the formation of cords(31) and Burbaud et al. have shown that disruption of the TPP synthesis, achieved through the mutation of the polyketide synthase gene (Pks), does not affect glycopeptidolipid (GPL) production(32), indicating that cord formation is dependent on several independent processes.

The whole *M. abscessus* genome was released In 2009(33), allowing for the detailed study of the molecular basis behind these phenotypic traits and virulence factors. Ripoll and colleagues sequenced the complete genome from *M. abscessus* ATCC 19977T strain (S phenotype), using a whole-genome shotgun strategy. Their findings show that its genome comprises a circular chromosome of 5,067,172 base pairs with 4,920 predicted coding sequences (EMBL accession number: CU458896) and also a mercury resistance plasmid of 23 kb (EMBL accession number: CU458745), similar to pMM23, found in *M. marinum*.

The authors reported a series of both mycobacterial and non-mycobacterial virulence factors, including the mammalian cell entry (MCE) operons, phospholipase C, and several antibiotic resistance genes. Especially interesting was the extensive amount of horizontally acquired genetic material, exhibiting synteny with a number of different genera, such as *Rhodococcus*, *Pseudomonas* and *Burkholderia*(33). These efforts have been curated in both general databases, like UniProt(34) (Proteome ID: UP000007137), and more specialized ones, like the Mycobrowser(35), which contains annotated genomic data from eight different mycobacterial species, including slow growers like *M. tuberculosis* and *M. leprae* and rapid growers, like *M. abscessus* and *M. smegmatis*.

However, structural information regarding *M. abscessus* proteins is scarce and, at the time of writing (07/02/2019), only 53 proteins, of which 39 are unique, had their structures determined, corresponding to less than 1% of the reported coding sequences.

Comprehensive understanding of protein function, the effects of mutations and successful structure-guided drug discovery is contingent on availability of accurate molecular models of the target proteins. These may be obtained either through experimental structure determination, or less costly, through computational modelling.

Our group has previously developed a database for the structural proteome of *M. tuberculosis*(36) using a set of in-house tools that rely on well-stablished databases, such as CATH(37) and SCOP(38) to retrieve domain-based information and use it in the modelling of the protein structures.

Besides *M. tuberculosis*, other organisms have had their structural proteomes modelled, many of which are available in the well-known UCSF ModBase(39) but, to our knowledge, such information is not available for *M. abscessus*. Additionally, online databases of protein models suffer from the insufficient computational sampling, since due to a throughput bottleneck, trade-offs are often made in what regards exhaustive structural sampling, an issue we tried to overcome with the approach proposed here.

In this work, we report the development of *Mabellini*, a novel, extensively annotated structural database, in which we focus on acquiring good-quality models for the *M. abscessus* proteome, providing an intuitive web interface, easily accessible to the potential end-users, who will be able to navigate through the proteomic data, and inspect and/or download the generated models. Knowledge concerning the protein structures is invaluable for understanding protein function and the effects caused by mutations. We also showcase examples of modelled proteins that are known to be implicated in some of the phenotypic features discussed above.

Methods

Protein structural profiles database (TOCCATA)

The modelling approach used in this work relies on a structural profile database (TOCCATA), which combines the domain-based information present in CATH(37) and SCOP(38) databases with information from primary data sources: The Protein Data Bank (PDB)(40) and UniProt(34). Each profile is annotated with functional information (derived from UniProt) and small-molecule-biding data (derived from PDB) and contains sets of pre-aligned representative structures, clustered by functional state. Each of the structures is annotated with the structural quality metrics, both directly derived from PDB (*e.g.* R-free, resolution) and those computed *e.g.* AEROSPACI(41,42) score. Due to incomplete annotation of the PDB in the AEROSPACI database, we fill in the missing values using a simple linear model, see Supplementary Material.

By identifying one profile from TOCCATA, our model-building pipeline (hereafter referred to as *Vivace*) accesses a large set of well-curated and annotated structures, permitting the identification of often non-obvious mappings between target sequences and template. TOCCATA contains three sets of profiles: single domain, multi domain, and *unclassified*. The first two are based on SCOP and CATH annotations, while the third contains all structures that are covered by neither SCOP nor CATH. Notably, sequences contained in each profile are clustered into groups sharing 95% identity (cluster representatives), in order to avoid including the same protein (or its variants) more than once, thus ensuring sufficient diversity of the profile.

A more detailed description of the TOCCATA architecture can be found in Ochoa-Montano(2015)(36).

Proteome modelling pipeline (Vivace)

Representative sequences for the proteome of *M. abscessus* retrieved from UniProt (Proteome ID: UP000007137) were organised as individual sequences, each of which formed a starting point for the pipeline. The full procedure for template selection, model generation and scoring are described below and depicted schematically in Figure 1. Thanks to *Vivace* being built using Ruffus(43), a pipelining solution, each run may progress at its own pace, allowing for optimal use of computational resources. This architecture allows for scaling the method to multiple machines, provided they share a common file system, allowing for use of *ad hoc* cluster solutions.

For each of the identified gene sequences (target sequences) that are longer than 250 residues, we perform domain decomposition, through a HMMER(44) search against PfamA(45) database. As much of TOCCATA is based on single domains, we ensure that we do query each of the putative domains against the profile database. If we were to rely only on the full-length sequence, domains with weaker evolutionary footprint (smaller, less conserved families) could remain undetected.

The full sequence and those of identified domains are subjected to PSI-BLAST searches against the UniRef100 database. The resulting sequences are re-aligned using MAFFT(46) and used for prediction of structural characteristics of the protein (e.g. secondary structure, solvent accessibility, etc.). This information in turn is fed into MELODY (part of the FUGUE(47) suite), a program that produces a compact representation of this information (profile), to be used for querying the TOCCATA database.

Using each of these profiles, Vivace performs FUGUE-based profile-profile search on *single*- and *multi-domain* subsets of the TOCCATA database. Only when no hit with sufficiently high Z-score (Z-score >= 8) is found, the *unclassified* set is searched. In so doing, we achieve a compromise between search speed and precision, as it can be safely assumed that a significant fraction of proteins in any given proteome comprises well-defined evolutionary domains, described in either SCOP or CATH. It is only when Vivace is unable to identify such domains, that a much more time-consuming search against the set of *unclassified* profiles is launched.

In parallel to the FUGUE search, Vivace conducts a BLAST search against pdbaa(48), a list of the protein sequences in PDB. The resulting hits with sequence identity of at least 70% are located in the TOCCATA database, and each profile that contains this protein or its subset is automatically included in the subsequent processing, even if it has not been detected by FUGUE search. In this way, we ensure that profiles containing very close homologs of the target are always included, even if there are already good hits present, identified from an initial FUGUE search.

Of the identified profiles, the ones with low confidence (FUGUE Z-Score below 4.0) are discarded. Profiles retrieved through BLAST searches have no inherent Z-Score associated and thus are always retained. All retained profiles undergo a grouping process, in order to reduce redundancy. The profiles are compared and are grouped where they cover an overlapping span of the queried sequences (over 75% overlap) and if they contain at least one SCOP or CATH family in common. Otherwise they form new *hit groups*. Each group is then further processed, and the query sequences are trimmed to match the length of the selected profiles, inside which the potential template structures reside. This step results on average in three hit groups per protein with 95% of the protein targets having at most seven hit groups (see Figure S1).

In each of the *hit groups*, Vivace retains only the profile clusters (cf. TOCCATA section) whose representative sequence presents no less than 20% identity to the query in comparison to the best one. However, if the cluster to be discarded affords a higher coverage of the query sequence and has at least 50% sequence identity, it is retained. This step permits pruning the template structures that may not be appropriate for successful modelling (preventing undue bias), while ensuring the highest possible coverage of the query by the templates.

In each of the selected cluster profiles, the potential templates are classified into functional states, with respect to the bound ligands and cofactors. There is also an *apo* state, denoting templates that are devoid of ligands, as well as *any* state, which contains all the templates available in the cluster. This step ensures, that each of the *states* comprises only structurally compatible, superimposable templates, making them well suited for modelling.

For each of the *states*, Vivace selects up to five templates for modelling. Initially, while there are still more then 5 candidates, the ones with low-coverage (less than two-thirds the maximum coverage) are iteratively removed. Then, the algorithm computes an all-against-all sequence identity matrix of all the templates in the *state*. The template set is then iteratively pruned, in a bottom-up approach, starting with the most mutually similar template pair and rejecting the one that presents the lower crystallographic quality (as measured by AEROSPACI scores present in TOCCATA database). This process is repeated until at most five templates per *state* remain, thus ensuring maximal diversity of template structures in terms of amino acid sequence, while maximizing their expected quality.

The selected templates in each of the *states* are re-aligned to the target sequence using BATON (manuscript in preparation), a modified open implementation of the COMPARER method(49) for protein structure alignment beyond rigid body superposition. This step allows slight adjustment to the pre-computed alignment present in TOCCATA, optimizing the fit between proteins involved in modelling, thus prospectively facilitating the subsequent model construction.

Given the BATON alignment of the selected templates in each of the states for each hit group in a protein, Vivace proceeds to the comparative modelling stage, using MODELLER(50) -- a macromolecular modelling suite, originally developed within the group. Using the *loopmodel* protocol implemented in MODELLER, we produce three models with two candidate loop realizations each. All the models are subject to the *very_fast* refinement protocol that is available in MODELLER suite. In our experience, increasing computational effort for refinement or producing greater number of models per state do not render substantially more useful models. Models are assessed with DOPE(51) and GA341(52) potentials.

Model Quality Assessment and selection

Modelling protocol renders tens (up to thousands, contingent on the number of available profile hits and states) of models per target, some of which are of substandard quality or very similar to the others. The experience of CHOPIN demonstrates that the end user prefers a limited number of models that help them explain the biological questions, rather than a comprehensive ensemble, which is often perceived as overwhelming. Thus, we have opted for displaying an ensemble of up to five models ranked according to their quality, as described below.

The first post-modelling processing stage consists of three different filtering steps for the removal of models exhibiting (i) extensive main chain clashes, (ii) large poorly resolved loops and (iii) loosely interacting ligands.

Then, to provide a meaningful model ensemble, we devised an algorithm to select models that: (i) are in a not yet displayed functional state (liganded, different conformation), (ii) cover a not yet displayed span of the protein or (iii) model a region of the protein with better accuracy than models already displayed.

To do so, we define two auxiliary quality metrics Q1 and Q2. The first metric (Q1) implicitly corresponds to global quality of the protein and is a linear combination of 80% structural consensus between models (computed by PconsD(53), via the comparison of residue-residue distance matrices, thus allowing for conformational flexibility in the comparison) and 20% stereochemical quality (computed by MolProbity,(54) in terms of MolProbity percentage score). The other metric (Q2) is a linear combination of GOAP-AG(55), GOAP-score and SOAP(56) potentials, scaled by reciprocal of the square of number of residues in the model. To avoid selecting incorrectly folded proteins and to discriminate between decoys with similar energy levels, we amend the Q2 metric with PconsD score at the same weight as each of the knowledge-based potentials (Figure S2)

Each of the constituent factors is expressed in terms of percentile across all the produced models. Therefore, a model with a Q1 or Q2 scores of 0.7 is explicitly better, in terms of this metric, than 70% of other models evaluated. High Q scores indicate that there is a significant difference in quality between models in the ensemble, and hence the high scoring models are much more likely to be correct, occupying low energy minimum for a given sequence. On the other hand – low scores do not necessarily indicate poor models, but rather models in an ensemble lacking structural consensus (in the case of Q1) or exhibiting relatively shallow free energy wells.

Gene Ontology (GO) annotation

Research groups are often interested in the biological functions of genes and in specific subsets thereof. Therefore, we constructed a clear, intuitive and interactive navigation tool that enables quick access to subsets of genes. For that, we implemented a *sunburst* representation of the GO terms with which *M. abscessus* proteins have been annotated.

Ontologies are sets of well-defined terms and relationships that represent the knowledge for and within a given domain. As such, the GO classification(57) actually relies on a complex set of relationships between the component terms, which do not form a simple hierarchical tree.

There are six basic relationships between GO terms: (i) *is_a*, (ii) *part_of*, (iii) *has_part*, (iv) *regulates*, (v) *positively_regulates* and (vi) *negatively_regulates*. Even considering solely the *is_a* relationship, the Gene Ontology is still more complex than a simple tree, in the sense that a child node may have more the one parent node. Therefore, the GO network can be defined as a Directed Acyclic Graph (DAG)(58), since there are no cycles, *i.e.* under no circumstance a node (GO term) can be a parent of itself. It is also unbalanced (different branches have variable depths) and ragged (there is actually no clear concept of "level" or "hierarchy"). Furthermore, it is worth noting that the GO terms are divided into three independent domains that are *not* connected by any “is a” relationship: (i) molecular function, (ii) biological process and (iii) cellular component(57).

We devised an algorithm (Hiera-GO, Figure 2, top box) that could use the parent-children relationships between the terms to create a simple hierarchical architecture. For this purpose, we used the *go_basic.obo* ontology, available at: http://geneontology.org/page/download-ontology, and parsed it to create a set of dictionaries, one for each domain (*i.e.* namespace) with the GO term as a key and the term attributes (*id, name, namespace, is_obsolete, is_a*) as an inner dictionary. We then used the *is_a* relationships to deduce the generations and to codify the terms, ultimately assigning a set of codes to each one of the terms, where each assigned code refers to a path through which that term can be accessed from the root node (Figure S3).

In a second step (Prote-GO, Figure 2, bottom box), the codified terms were used to construct a hierarchical representation of *M. abscessus* proteome annotation retrieved from Uniprot. Subsequently, this hierarchical representation was written to a JSON file, which was adopted in the construction of the interactive sunburst. Since the genes are conventionally annotated with only the most specific GO term, we also created a set of functions to annotate them with all their parent terms, in order to be able to query the relevant subset by using more general terms that encompass their functions. Both programs are available as supplementary files.

Mabellini Database

In addition to the modelled structures and functional annotation (e.g. GO terms), Mabellini data are mapped to both Uniprot and Pfam and also track the predicted protein properties (including membrane protein topology and disorder). Membrane protein topology has been predicted with TOPCONS(59) metapredictor, including the signal peptide predictor as implemented in SPOCTOPUS.(60) Disordered regions in the evaluated proteins have been assessed with DISOPRED.(61) Both approaches were chosen for their robustness and prevalence in the field, as well as capability for a relatively high throughput.

The database itself has been deployed on a PostgreSQL server and was implemented using SQLAlchemy Python module. The database schema (Figure S4) follows Codd's third normal form (3NF), thus making for a more robust and future-safe representation.

Website development

The website back-end is managed using the SQLAlchemy Python module, which allows for the direct access and query of the underlying database populated by Vivace. In order to decrease query retrieval time, and thus allow more users to use Mabellini DB at a time, we parallelized all the tasks using the Multiprocessing Python module. Furthermore, using a combination of publicly available Application Programming Interfaces (APIs) and scripts developed in-house, we provide users direct mappings to other well-established database resources such as: PDB, Uniprot, Pfam, SCOP and CATH.

The website is extensively based on the Flask module, which is used to generate the website pages and provide connectivity with the SQL database. The website is written in HTML5 using CSS, Javascript and JQuery as well as the Bootstrap (version 4) framework. JINJA2 templating language for Python was used to dynamically generate HTML templates. The sunburst charts were created using the D3.js package, using the information generated by Hiera-Go and Prote-GO algorithms in JSON format as data source.

In order to facilitate programmatic usage of the Mabellini database, significant fraction of data stored there in has been made available through a RESTful API. Data is returned in form of JSON objects, thus is easily parsable by downstream software and straightforwardly incorporated into existing solutions.

Results

Structural proteome characteristics

We have built a comprehensive set of *M. abscessus* protein structures, covering 3,394 out of 4,940 protein coding sequences annotated in the UniProt UP000007137 proteome. This comprises 13,879 models, of which 1,855 are bound to a ligand (cofactors, substrates, small molecules, etc. which are retrieved from the structures used as templates). Despite the paucity of experimentally determined protein structures of *M. abscessus,* with only 53 entries in the PDB, Mabellini now addresses 68.7% of gene products.

The models included in the final database had a median length of 211 residues, with the longest (MAB_2301c: Rank 01, a putative mmpL transporter), being 946 residues long. The average gene coverage of our models is 75.4% while the median gene coverage is 92.1%, showing that we have mostly generated near-complete proteins. The average Molprobity score was of 3.21 with a standard deviation of 0.42, when considering all models conjointly, but it is lower (3.15) when considering only non-liganded models and higher (3.56) when considering only the liganded models, likely due to the algorithm employed for the incorporation of ligands, as described in the methods section. The Molprobity score is a log-weighted value that takes into consideration atomic clashes, main-chain and side-chain features, and the resulting number represents the crystallographic resolution at which those features would be expected(54).

The average identity of the modelled protein sequences to the most similar template was of 31.7%, with a first quartile value of 18.2% and third quartile of 38%, yielding an interquartile range of 19.8%. As our approach is profile-based, some of the templates had identities significantly below the generally established limits for comparative modelling, and yet rendered models with structural features compatible with evolutionary domain annotations. Further statistical information regarding the models' properties may be found in Figure 3.

Every produced model has been subjected to a multi-faceted quality assessment, using approaches described in the Methods section. Mabellini makes all of the produced models available for download, together with the quality assessment results; each of the models available in the website has the quality metrics embedded into the PDB file header, as well as a local quality estimate in the B-factor column.

Modelled virulence factors

To illustrate some of the results generated by our pipeline, we selected a set of biologically-relevant proteins involved either in important metabolic pathways or in antibiotic resistance. In order to be able to comment on the most likely functional state of a given target, we performed a simple PSI-Blast search to infer its oligomeric state and selected, for the following analysis, proteins that are likely functional in a monomeric state.

Table 1 summarizes the proteins presented in this section together with the source of evidence for their biological relevance and features of the best-ranking model generated for them.

Proteins MAB_0591 (Figure 4A), MAB_1496 (Figure 4B) and MAB_2385 (Figure 4C) have been recently reviewed in (22) and are involved with resistance to the antibiotics rifamycin, tetracycline and aminoglycosides, respectively.

MAB_0591 (Arr_Mab) is an enzyme that catalyses the ADP-ribosylation of rifamycin antibiotics, thus conferring resistance to this class of compounds. 137 of the 141 residues in this enzyme were modelled, yielding a model with a Molprobity score of 2.861 and high structural similarity to the used template from *M. smegmatis* (CATH ID: 2hw2A00, TM-score: 0.99, RMSD: 0.09 Å). We were able to model the liganded state, in which it is complexed to Rifampicin. As this model is built based on a very close homolog, it has a high expected verisimilitude, which means it could plausibly serve as a target in molecular docking (and re-docking) studies.

MAB_1496c encodes a flavoprotein monooxygenase implicated in resistance to tetracyclines. Interestingly, among the flavoprotein templates selected, a rifampicin monooxygenase from *N. farcinica* (PDB ID: 5KOX) shares 31.9% identity with the M. abscessus sequence. The top-ranked model is bound to the FAD cofactor. Superimposition of all the templates onto the Rank 01 model yields a maximum RMSD of 0.86 Å, which indicates a largely structurally conserved protein, considering that the sequence identities range from 29.9 to 32.4 %. When compared to all templates, MAB_1496 presents a 12-residue-long insertion at position 76, resulting in a longer loop formed between helices 3 and 4 (Figure 4B, black arrowhead), near the substrate-binding site, which could be related to interaction specificity.

MAB_2385 codes for an 3"-O-phosphotranferase, which transfers a phosphate group from an ATP molecule to the N-methyl-L-glucosamine moiety of streptomycin and is implicated in resistance to that compound. All models generated for this protein have coverage of 99.6% and the best models' RMSD values to the templates range from 0.42 to 0.86 Å. The most notable difference between the model and templates is the absence of a C-terminal α-helix in our model (Figure 4C, black arrowhead).

The protein Erm(41), encoded by the MAB_2297 gene, is one of the main effectors for clarithromycin resistance. Despite having low sequence Identity to the templates used in the modelling (17.3 - 23.3%), the generated models cover 93.6 - 99.6% of the protein length, and exhibit a correct topology, bearing a canonical Rossman fold domain. The Rank 01 model is bound to S-adenosylmethionine (SAM, AdoMet), which is a cofactor for class I methyltransferases. All models contain the characteristic nucleotide-binding GxGxG motif(62) at the end of the first β-strand and the acidic residue GLU60 at the end of the second β-strand, which interacts with the SAM molecule via the hydroxylic groups of the ribose moiety (Figure 4D, in detail). The templates employed in the construction of the Rank 01 model bear an extra helical C-terminal domain, which is likely related to substrate specificity,(63) and is lacking in MAB_2297 gene (Figure 4D, black arrowhead).

Besides antibiotic resistance, we have also chosen to use as examples proteins involved in biosynthetic pathways that determine some important phenotypic aspects, such as the production of GPL and TPP.

Recently, TPP was suggested as the likely candidate for the formation of cord-like structures in *M. abscessus* and a specific genomic locus, which comprises the genes MAB_0934, MAB_0935c, MAB_0936c, MAB_0937c, MAB_0938c and MAB_0939, was identified as being responsible for the TPP biosynthesis(32).

MAB_0935c codes for the FadD protein, an acyl-CoA syntethase, which is thought to activate and transfer C16 saturated fatty acids to the polyketide synthase (Pks) enzyme MAB_0939. Our best model for the MAB_0935c protein used PDB 5D6J as template, which is the FadD32 from *M. smegmatis*. The most noticeable difference between the two proteins is a large insertion in FadD32 from *M. smegmatis* when compared to MAB_0935c, which occurs between THR491 and GLU492 of our model (Figure 4E, black arrowhead). Nevertheless, the RMSD between the two structures was of 0.28 Å.

Also involved in the production of TPP is the PapA3 protein (Figure 4F), encoded by MAB_0938c, whose function consists of transferring the fatty acyl group elongated by Pks onto trehalose. Despite very low sequence identity (15.3%), the TM-Score between the best model and the 5T3E template was of 0.94, indicating an excellent superimposition of the two structures. Both MAB_0938 and the 5T3E PDB template have been annotated with the Pfam domain PF00668, thus increasing the confidence of the inferred fold topology.

Another important genomic feature described in (64) is the GPL locus, comprising genes that range from MAB_4097 to MAB_4117. We were able to generate models for 11 out of the 21 genes present in this locus (two of which are putative proteins): MbtH-like protein (MAB_4100c), Fmt (MAB_4103c), Rmt3 (MAB_4105c), Rmt4 (MAB_4108c), Rmt2 (MAB_4109c), RmlB (MAB_4111c), Gtf3 (MAB_4112c), RmlA (MAB_4113), MmpL4b (MAB_4115c), MmpL4a (MAB_4116c) and MMpS4 (MAB_4117c) with coverages for the best models in each case ranging from 33.5% to 99.7%.

Web-page front end

The website is accessible at the following URL: https://www.mabellinidb.science. For simplicity, the first page presents all types of queries available to the user (Figure 5A). By selecting the appropriate tag (Figure 5A top panel), the user has the option to search directly the *Mycobacterium abscessus* proteome by: gene id, Uniprot id, Pfam id, sequence similarity, Gene Onthology (GO) id and Enzyme Commission id (EC). However, the end-user may not have prior knowledge of such terms, hence for each search field we implemented an additional text (keyword) search. For example, the RNA polymerase-binding protein (RbpA) has the Uniprot ID: 'B1MAL8'; to retrieve all the genes with this Uniprot ID one can either directly input the exact id or use terms such as: 'RNA polymerase', 'pol' or 'RNA'.

Furthermore, to make it even more straightforward for the end user to navigate the *Mycobacterium abscessus* proteome, we created four sunburst queries (Figure 5A, bottom panel) based on GO terms and EC numbers. Figure 5B presents the sunburst for Biological Process. Here, users can click and hover over each segment of the sunburst. For example, if one is interested in genes involved in *alcohol metabolic processes,* he/she will first select *metabolic process*, then *small molecule metabolic process* and finally *alcohol metabolic process*.

This allows the user to intuitively explore the hierarchical architecture we implemented. Moreover, at the top of the page, a breadcrumb trail (Figure 5B) will present each selection (click or hover) to inform the user about his/her position on the sunburst.

Once the desired level is achieved the user can click the *Query: term* button which will lead to the Multiple Results page (Figure 5C). In this page, the user is provided with a fully searchable summary table presenting all the hits as well as the query terms used, the total number of genes retrieved and a ‘Download’ button. In the summary table for each entry we provide the gene's id (ordered locus), name, length, the percentage modelled by our algorithm, its Uniprot and Pfam ids, the EC codes (when available) and the number of models generated. The genes for which no models were produced (due to the lack of suitable templates) are highlighted in red in the query results table. For all other entries, the user can access the models by selecting (clicking) the gene id of interest.

The Models page (Figure 6) is divided into three sections. At the top we present background information about the gene of interest, as well as mappings to Pfam domains and EC classification (Figure 6A, at the top). In the following section we display a list of generated models split in two tables that reflect the model's liganded state (Figure 6A, at the bottom). The tables contain the models' qualities, coverage, length, residue interval and the templates used in the modelling process. If there are experimental structures available, they are also displayed as an additional table at the top of this section. The last section presents an interactive panel, built using NGL, where the user can inspect any model selected in the tables presented above (Figure 4B). For example, using the Model Controls the user can display the surface contour, colour by different properties or according to the Pfam domains, view the ligand interactions, save an image file, etc. Additionally, the user can download the structures by clicking the *Download* (for individual entries) or *Download all models* (for batch download) buttons.

The implementation of a RESTful JSON-based API in Mabellini increases its utility for the advanced user intending to incorporate the database in their software solution. Through a set of powerful, yet simple, query mechanisms (see

Supplementary Material and description in the website), one may selectively query the database and retrieve the desired subset of data stored therein. Due to a low-level implementation, sharing the query engine with the website, this API is both fast and robust, permitting for a substantial traffic and self-limiting in cases of excessive usage. Even though website and API share the same engine, they are mutually independent, thus performance of one is not fully contingent on the other.

In summary, we have created an intuitive and functional website that should be of interest both to researchers interested in the biological aspects of *Mycobacterium abscessus* and those generally interested in protein structure prediction and proteome modelling.

Discussion

Genomic and proteomic databases are becoming increasingly widespread. Presently, there are multiple online services available, such as the GenBank(65), UniProt, KEGG(66) and Ensembl(67) databases, in which it is possible to find extensive information regarding the gene products. Bacteria-specific databases have also been developed in the past decades, as is the case of BacMap(68), a genome atlas where one can explore the complete genome of over 1,790 bacterial species. These more specific databases usually have a more direct approach in the data presentation, normally best suited to the target group of users. The Mycobrowser service, for example, contains genomic information for eight different mycobacterial species, including *M. leprae*, *M. tuberculosis* and *M. abscessus*. It provides tools to explore the whole-genomes of these microorganisms, including summarized genome data and the sequences of individual genes. Nevertheless, these databases are usually dedicated to the curation and displaying of protein and gene sequences, and few of them tackle the three-dimensional structure of proteins.

Some of the most recent resources, such as the Genome3D(69,70), ModBase and even the COSMIC(71) database (the Catalogue of Somatic Mutations in Cancer), have invested in the three-dimensional representation of proteomic data. Genome3D is a recently-established collaborative database which aims to coalesce structure and sequence-based annotations, as well as three-dimensional models, in a centralized hub, providing consensus annotations for a number of model organisms (currently ten different organisms, including *H. sapiens, C. elegans, E. coli S. cerevisiae,* and *D. melanogaster*). Progress is underway to integrate the *Vivace* pipeline as an additional modelling source.

Chopin, the *M. tuberculosis*-related resource based on *Vivace* pipeline, released by our group in 2015, has attracted interest. Since then, we have rectified its observed shortcomings in terms of template availability, template selection and model selection. We have updated our underlying database (TOCCATA) to reflect the current state of knowledge and instituted protocols for nearly fully automated updates in the future. In comparison to the TOCCATA version used to build the original CHOPIN, the new version contains nearly twice the amount of data, with 10842 single-domain profiles and 4162 multi-domain ones, comprising over 604,000 structural templates (PDBs), while the older version contained 6590 and 2261 single- and multi-domain profiles and approximately 314,000 structures. Additionally, the new TOCCATA contains 12,103 “unclassified” profiles, containing structures with no SCOP/CATH domain assignments, to be searched in case of lack of satisfactory hit among the single- and multi-domain profile sets. We have also optimized the parameters for the template selection, allowing for a higher expected quality of predictions. Finally, we have also conducted a sanity check and enforced modelling using the "unclassified" profiles that contained structures identified by a BLAST search against PDB.

Since our approach is mainly based on comparative modelling, although we have been able to generate models for roughly 75% of the M. abscessus proteome, we are still restricted by structural data available in public repositories for close homologues. In our experience, structural information about complex multi-domain gene products is especially scarce, a tendency that is likely to change as a result of the recent advances in alternative methods for protein structure determination, such as cryo-EM. Ab-initio methods for protein modelling, on the other hand, are still not reliable enough to be incorporated in databases like Mabellini, since the usefulness of the resulting models would questionable. Therefore, endeavours such as the NIH-funded Protein Structure Initiative and the several structural genomics projects(72) remain of paramount importance in enabling the generation of high-confidence comparative models.

Furthermore, Mabellini is based on TOCCATA, which is inherently oriented toward evolutionary domains. This yields models that often do not span the entire protein. While for many proteins we have produced well-substantiated models of multi-domain proteins based on multi-domain homologues, there is still room for improvement where there is no homologue. We intend to address this aspect in the upcoming updates to the methodology and to the resource itself.

The in-depth analysis of some biologically relevant targets showed that the models used to populate Mabellini exhibit a correct fold and satisfactory stereochemistry, thus being highly valuable. They may serve a number of distinct purposes that include, but are not restricted to: general structure investigation, understanding function and selectivity, target fishing, molecular docking, virtual screening, molecular dynamics and the analysis of the impacts of mutations.

Conclusion

Mabellini, the comprehensive resource for theoretical models of *Mycobacterium abscessus* proteins is, to the best of our knowledge, the first resource of its kind for this particular pathogen. While species-specific databases for bacteria have been developed for decades, they predominantly focus on genomic characteristics, instead of structure and function.

The methodology used in the construction of Mabellini is a major advance over the one used for Chopin. While the underlying rationale remains largely the same, we have improved homology detection and introduced multi-domain and full-chain ("unclassified") profiles into database. The "fast track" link between PDB and TOCCATA (through an initial BLAST search) ensures we do not miss any of the close-to-native structures.

We have built an entirely new, streamlined web-based user interface for Mabellini, which allows for an easy access to the data, with special focus on usability and ease of access to information.

Besides performing regular updates, our goal is to also aggregate to the database functional annotations and protein-protein interactions derived from publicly available sources, pathway analysis (using databases such as BioCyc(73)), assessment of effects of mutations through state-of-art methods developed by our group (SDM(74) and mCSM(75) family of methods), as well as genome-wide coupling analysis results(76). Efforts are underway towards the development of robust methods to model accurately the structure of multimeric proteins. We want to move beyond simple comparative modelling to allow for constructing and accurately selecting correct assembly modes.

Acknowledgements

We thank the following funding agencies for providing support to the Blundell and Floto groups. MJS: The Botnar Foundation (Project 6063 - RG91317); PHMT: The Cystic Fibrosis Trust (SRC 010 - RG92232); BB: the UK Medical Research Council (MRC-DBT Centre Partnership); LC: Ipsen (Cambridge Studentship -08032016). The authors also thank Vitor Mendes, Sherine Thomas, Sheikh Mohammed Arif, Marta Acebron, Sundeep Chaitanya, Asma Munir and Michal Blaszczyk for helpful discussions.

References

1. Tortoli, E., Kohl, T. A., Brown-Elliott, B. A., et al. (2016) Emended description of mycobacterium abscessus mycobacterium abscessus subsp. Abscessus and mycobacterium abscessus subsp. bolletii and designation of mycobacterium abscessus subsp. massiliense comb. nov., *Int. J. Syst. Evol. Microbiol.*, **66**, 4471–4479.

2. Sassi, M. and Drancourt, M. (2014) Genome analysis reveals three genomospecies in Mycobacterium abscessus, 1–10.

3. Adékambi, T., Reynaud-gaubert, M., Greub, G., et al. (2004) Amoebal Coculture of Mycobacterium massiliense sp . nov . from the Sputum of a Patient with Hemoptoic Pneumonia, 5493–5501.

4. Adékambi, T., Berger, P., Raoult, D., et al. (2006) rpoB gene sequence-based characterization of emerging non-tuberculous mycobacteria with descriptions of Mycobacterium bolletii sp. nov., Mycobacterium phocaicum sp. nov. and Mycobacterium aubagnense sp. nov, *Int. J. Syst. Evol. Microbiol.*, **56**, 133–143.

5. Drancourt, M. (2014) Looking in amoebae as a source of mycobacteria, *Microb. Pathog.*, **77**, 119–124.

6. Medjahed, H., Gaillard, J. L. and Reyrat, J. M. (2010) Mycobacterium abscessus: a new player in the mycobacterial field, *Trends Microbiol.*, **18**, 117–123.

7. Esther, C. R., Esserman, D. A., Gilligan, P., et al. (2010) Chronic Mycobacterium abscessus infection and lung function decline in cystic fibrosis, *J. Cyst. Fibros.*, **9**, 117–123.

8. Qvist, T., Pressler, T., Thomsen, V. O., et al. (2013) Nontuberculous mycobacterial disease is not a contraindication to lung transplantation in patients with cystic fibrosis: A retrospective analysis in a Danish patient population, *Transplant. Proc.*, **45**, 342–345.

9. Gilljam, M., Scherstén, H., Silverborn, M., et al. (2010) Lung transplantation in patients with cystic fibrosis and Mycobacterium abscessus infection, *J. Cyst. Fibros.*, **9**, 272–276.

10. Bryant, J. M., Grogono, D. M., Greaves, D., et al. (2013) Whole-genome sequencing to identify transmission of Mycobacterium abscessus between patients with cystic fibrosis: A retrospective cohort study, *Lancet*, **381**, 1551–1560.

11. Bryant, J. M., Grogono, D. M., Rodriguez-Rincon, D., et al. (2016) Emergence and spread of a human-transmissible multidrug-resistant nontuberculous mycobacterium, *Science (80-.).*, **354**, 751–757.

12. Sander, P., Prammananan, T., Onyi, G. O., et al. (2008) A Single 16S Ribosomal RNA Substitution Is Responsible for Resistance to Amikacin and Other 2-Deoxystreptamine Aminoglycosides in Mycobacterium abscessus and Mycobacterium chelonae, *J. Infect. Dis.*, **177**, 1573–1581.

13. Nessar, R., Reyrat, J. M., Murray, A., et al. (2011) Genetic analysis of new 16s rRNA mutations conferring aminoglycoside resistance in Mycobacterium abscessus, *J.*

Antimicrob. Chemother., **66**, 1719–1724.

14. Nash, K. A., Brown-Elliott, A. B. and Wallace, R. J. (2009) A Novel gene, erm(41), confers inducible macrolide resistance to clinical isolates of mycobacterium abscessus but is absent from mycobacterium chelonae, *Antimicrob. Agents Chemother.*, **53**, 1367–1376.

15. Maurer, F. P., Castelberg, C., Quiblier, C., et al. (2014) Erm(41)-dependent inducible resistance to azithromycin and clarithromycin in clinical isolates of mycobacterium abscessus, *J. Antimicrob. Chemother.*, **69**, 1559–1563.

16. Qvist, T., Pressler, T., Høiby, N., et al. (2014) Shifting paradigms of nontuberculous mycobacteria in cystic fibrosis, *Respir. Res.*, **15**, 1–6.

17. Floto, R. A., Olivier, K. N., Saiman, L., et al. (2016) US Cystic Fibrosis Foundation and European Cystic Fibrosis Society consensus recommendations for the management of non-tuberculous mycobacteria in individuals with cystic fibrosis, *Thorax*, **71**, i1–i22.

18. Jeon, K., Kwon, O. J., Lee, N. Y., et al. (2009) Antibiotic Treatment of *Mycobacterium abscessus* Lung Disease, *Am. J. Respir. Crit. Care Med.*, **180**, 896–902.

19. Wallace-Jr, R. J., Meier, A., Brown, B. A., et al. (1996) Genetic basis for clarithromycin resistance among isolates of Mycobacterium chelonae and Mycobacterium abscessus . These include : Genetic Basis for Clarithromycin Resistance among Isolates of Mycobacterium chelonae and Mycobacterium abscessus, **40**, 1676–1681.

20. Takehiko Kobayashi, Tsuyuguchi, K., Yoshida, S., et al. (2017) Mycobacterium abscessus subsp. abscessus Lung Disease: Drug Susceptibility Testing in Sputum Culture Negative Conversion, *Int. J. Mycobacteriology*, **6**, 239–245.

21. Koh, W.-J., Jeong, B.-H., Kim, S.-Y., et al. (2017) Mycobacterial Characteristics and Treatment Outcomes in *Mycobacterium abscessus* Lung Disease, *Clin. Infect. Dis.*, **64**, 309–316.

22. Luthra, S., Rominski, A. and Sander, P. (2018) The role of antibiotic-target-modifying and antibiotic-modifying enzymes in mycobacterium abscessusdrug resistance, *Front. Microbiol.*, **9**, 1–13.

23. Rüger, K., Hampel, A., Billig, S., et al. (2014) Characterization of rough and smooth morphotypes of mycobacterium abscessus isolates from clinical specimens, *J. Clin. Microbiol.*, **52**, 244–250.

24. Howard, S. T., Rhoades, E., Recht, J., et al. (2006) Spontaneous reversion of Mycobacterium abscessus from a smooth to a rough morphotype is associated with reduced expression of glycopeptidolipid and reacquisition of an invasive phenotype, *Microbiology*, **152**, 1581–1590.

25. Byrd, T. F. and Lyons, C. R. (1999) Preliminary characterization of a Mycobacterium abscessus mutant in human and murine models of infection., *Infect. Immun.*, **67**, 4700–7.

26. Park, I. K., Hsu, A. P., Tettelin, H., et al. (2015) Clonal diversification and changes in lipid traits and colony morphology in mycobacterium abscessus clinical isolates, *J. Clin. Microbiol.*, **53**, 3438–3447.

27. Sánchez-Chardi, A., Olivares, F., Byrd, T. F., et al. (2011) Demonstration of cord formation by rough Mycobacterium abscessus variants: Implications for the clinical microbiology laboratory, *J. Clin. Microbiol.*, **49**, 2293–2295.

28. Julián, E., Roldán, M., Sánchez-Chardi, A., et al. (2010) Microscopic cords, a virulence-related characteristic of Mycobacterium tuberculosis, are also present in nonpathogenic mycobacteria, *J. Bacteriol.*, **192**, 1751–1760.

29. Bernut, A., Herrmann, J.-L., Kissa, K., et al. (2014) Mycobacterium abscessus cording prevents phagocytosis and promotes abscess formation, *Proc. Natl. Acad. Sci.*, **111**, E943–E952.

30. Brambilla, C., Llorens-Fons, M., Juliï¿½n, E., et al. (2016) Mycobacteria Clumping Increase Their Capacity to Damage Macrophages, *Front. Microbiol.*, **7**, 1–12.

31. Llorens-Fons, M., Pérez-Trujillo, M., Julián, E., et al. (2017) Trehalose polyphleates, external cell wall lipids in Mycobacterium abscessus, are associated with the formation of clumps with cording morphology, which have been associated with virulence, *Front. Microbiol.*, **8**, 1–15.

32. Burbaud, S., Laval, F., Lemassu, A., et al. (2016) Trehalose Polyphleates Are Produced by a Glycolipid Biosynthetic Pathway Conserved across Phylogenetically Distant Mycobacteria, *Cell Chem. Biol.*, **23**, 278–289.

33. Ripoll, F., Pasek, S., Schenowitz, C., et al. (2009) Non mycobacterial virulence genes in the genome of the emerging pathogen Mycobacterium abscessus, *PLoS One*, **4**.

34. Bateman, A., Martin, M. J., O'Donovan, C., et al. (2017) UniProt: The universal protein knowledgebase, *Nucleic Acids Res.*, **45**, D158–D169.

35. Kapopoulou, A., Lew, J. M. and Cole, S. T. (2011) The MycoBrowser portal: A comprehensive and manually annotated resource for mycobacterial genomes, *Tuberculosis*, **91**, 8–13.

36. Ochoa-montano, B., Mohan, N., Blundell, T. L., et al. (2015) CHOPIN: a web resource for the structural and functional proteome of Mycobacterium tuberculosis, *Database (Oxford).*, **2015**, 1–10.

37. Sillitoe, I., Lewis, T. E., Cuff, A., et al. (2015) CATH: Comprehensive structural and functional annotations for genome sequences, *Nucleic Acids Res.*, **43**, D376–D381.

38. Andreeva, A., Howorth, D., Chothia, C., et al. (2014) SCOP2 prototype: A new approach to protein structure mining, *Nucleic Acids Res.*, **42**, 310–314.

39. Pieper, U., Webb, B. M., Dong, G. Q., et al. (2014) ModBase, a database of annotated comparative protein structure models and associated resources, *Nucleic Acids Res.*, **42**, 1–11.

40. Burley, S. K., Berman, H. M., Bhikadiya, C., et al. (2019) Protein Data Bank: the single global archive for 3D macromolecular structure data, *Nucleic Acids Res.*, **47**, D520–D528.

41. Fox, N. K., Brenner, S. E. and Chandonia, J. M. (2014) SCOPe: Structural Classification of Proteins - Extended, integrating SCOP and ASTRAL data and classification of new structures, *Nucleic Acids Res.*, **42**, 304–309.

42. Chandonia, J.-M. (2004) The ASTRAL Compendium in 2004, *Nucleic Acids Res.*, **32**, 189D – 192.

43. Goodstadt, L. (2010) Ruffus: A lightweight python library for computational pipelines, *Bioinformatics*, **26**, 2778–2779.

44. Eddy, S. R. (1995) Multiple alignment using hidden Markov models., *Proceedings. Int. Conf. Intell. Syst. Mol. Biol.*, **3**, 114–20.

45. Finn, R. D., Bateman, A., Clements, J., et al. (2014) Pfam: The protein families database, *Nucleic Acids Res.*, **42**, 222–230.

46. Katoh, K., Misawa, K., Kuma, K., et al. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform., *Nucleic Acids Res.*, **30**, 3059–3066.

47. Shi, J., Blundell, T. L. and Mizuguchi, K. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties., *J. Mol. Biol.*, **310**, 243–257.

48. Wang, G. and Dunbrack, R. L. (2003) PISCES: a protein sequence culling server, *Bioinformatics*, **19**, 1589–1591.

49. Šali, A. and Blundell, T. L. (1990) Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming, *J. Mol. Biol.*, **212**, 403–428.

50. Šali, A. and Blundell, T. L. (1993) Comparative Protein Modelling by Satisfaction of Spatial Restraints, *J. Mol. Biol.*, **234**, 779–815.

51. Melo, F., Sánchez, R. and Sali, A. (2009) Statistical potentials for fold assessment, *Protein Sci.*, **11**, 430–448.

52. Melo, F. (2007) Fold assessment for comparative protein structure modeling, *Protein Sci.*, **16**, 2412–2426.

53. Skwark, M. J. and Elofsson, A. (2013) PconsD: Ultra rapid, accurate model quality assessment for protein structure prediction, *Bioinformatics*, **29**, 1817–1818.

54. Chen, V. B., Arendall, W. B., Headd, J. J., et al. (2010) MolProbity: All-atom structure validation for macromolecular crystallography, *Acta Crystallogr. Sect. D Biol. Crystallogr.*, **66**, 12–21.

55. Zhou, H. and Skolnick, J. (2011) GOAP: A generalized orientation-dependent, all-atom statistical potential for protein structure prediction, *Biophys. J.*, **101**, 2043–2052.

56. Dong, G. Q., Fan, H., Schneidman-Duhovny, D., et al. (2013) Optimized atomic statistical potentials: Assessment of protein interfaces and loops, *Bioinformatics*, **29**, 3158–3166.

57. Ashburner, M., Ball, C. A., Blake, J. A., et al. (2000) Gene ontology: Tool for the unification of biology, *Nat. Genet.*, **25**, 25–29.

58. Mostafavi, S. and Morris, Q. (2009) Using the Gene Ontology Hierarchy when Predicting Gene Function, *'09 Proc. Twenty-Fifth Conf. Uncertain. Artif. Intell.*, **cs.LG**, 419–427.

59. Tsirigos, K. D., Peters, C., Shu, N., et al. (2015) The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides, **43**, 401–407.

60. Viklund, H., Bernsel, A., Skwark, M., et al. (2008) SPOCTOPUS : a combined predictor of signal peptides and, **24**, 2928–2929.

61. Ward, J. J., Sodhi, J. S., Mcguffin, L. J., et al. (2004) Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life, 635–645.

62. Schubert, H. L., Blumenthal, R. M. and Cheng, X. (2003) Many paths to methyltransfer: a chronicle of convergence, *Trends Biochem. Sci.*, **28**, 329–335.

63. Martin, J. (2002) SAM (dependent) I AM: the S-adenosylmethionine-dependent methyltransferase fold, *Curr. Opin. Struct. Biol.*, **12**, 783–793.

64. Ripoll, F., Deshayes, C., Pasek, S., et al. (2007) Genomics of glycopeptidolipid biosynthesis in Mycobacterium abscessus and M. chelonae, *BMC Genomics*, **8**, 1–9.

65. Benson, D. A., Cavanaugh, M., Clark, K., et al. (2013) GenBank, *Nucleic Acids Res.*, **41**, 36–42.

66. Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto encyclopedia of genes and genomes, *Nucleic Acids Res.*, **28**, 27–30.

67. Zerbino, D. R., Achuthan, P., Akanni, W., et al. (2018) Ensembl 2018, *Nucleic Acids Res.*, **46**, D754–D761.

68. Cruz, J., Liu, Y., Liang, Y., et al. (2012) BacMap: An up-to-date electronic atlas of annotated bacterial genomes, *Nucleic Acids Res.*, **40**, 599–604.

69. Lewis, T. E., Sillitoe, I., Andreeva, A., et al. (2013) Genome3D: A UK collaborative project to annotate genomic sequences with predicted 3D structures based on SCOP and CATH domains, *Nucleic Acids Res.*, **41**, 499–507.

70. Lewis, T. E., Sillitoe, I., Andreeva, A., et al. (2015) Genome3D: Exploiting structure to help users understand their sequences, *Nucleic Acids Res.*, **43**, D382–D386.

71. Forbes, S. A., Beare, D., Boutselakis, H., et al. (2017) COSMIC: Somatic cancer genetics at high-resolution, *Nucleic Acids Res.*, **45**, D777–D783.

72. Ewing, R. M. and Doyle, D. A. (2015) Structural Proteomics: Large-Scale Studies, *eLS*, 1–7.

73. Karp, P. D., Billington, R., Caspi, R., et al. (2017) The BioCyc collection of microbial genomes and metabolic pathways, *Brief. Bioinform.*, 1–9.

74. Worth, C. L., Preissner, R. and Blundell, T. L. (2011) SDM - A server for predicting effects of mutations on protein stability and malfunction, *Nucleic Acids Res.*, **39**, 215–222.

75. Pires, D. E. V, Ascher, D. B. and Blundell, T. L. (2014) MCSM: Predicting the effects of mutations in proteins using graph-based signatures, *Bioinformatics*, **30**, 335–342.

76. Skwark, M. J., Croucher, N. J., Puranen, S., et al. (2017) Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis

analysis, *PLOS Genet.*, **13**, e1006508.

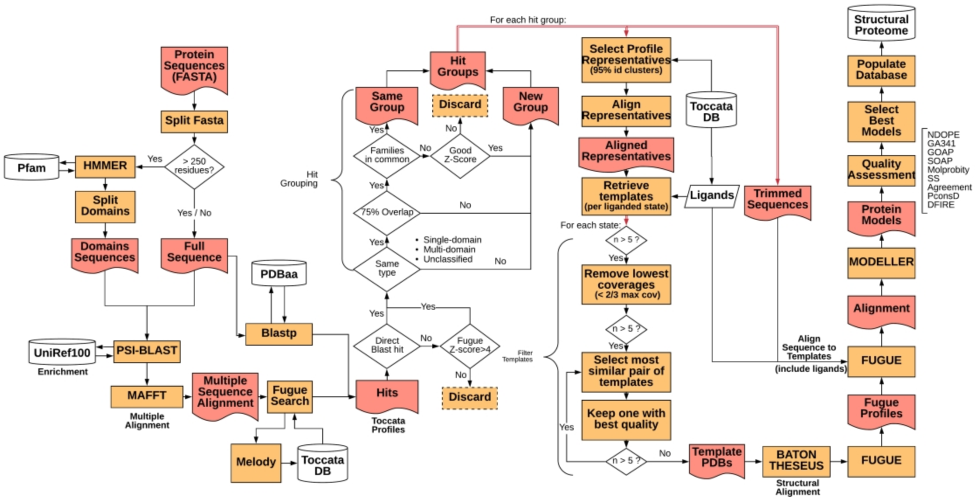


Figure 1 - Simplified flowchart of Vivace pipeline execution. Orange boxes represent processes, red nodes represent generated files, white cylinders represent databases and white diamonds represent decision-making nodes. The two red double arrows denote branching steps of the pipeline.

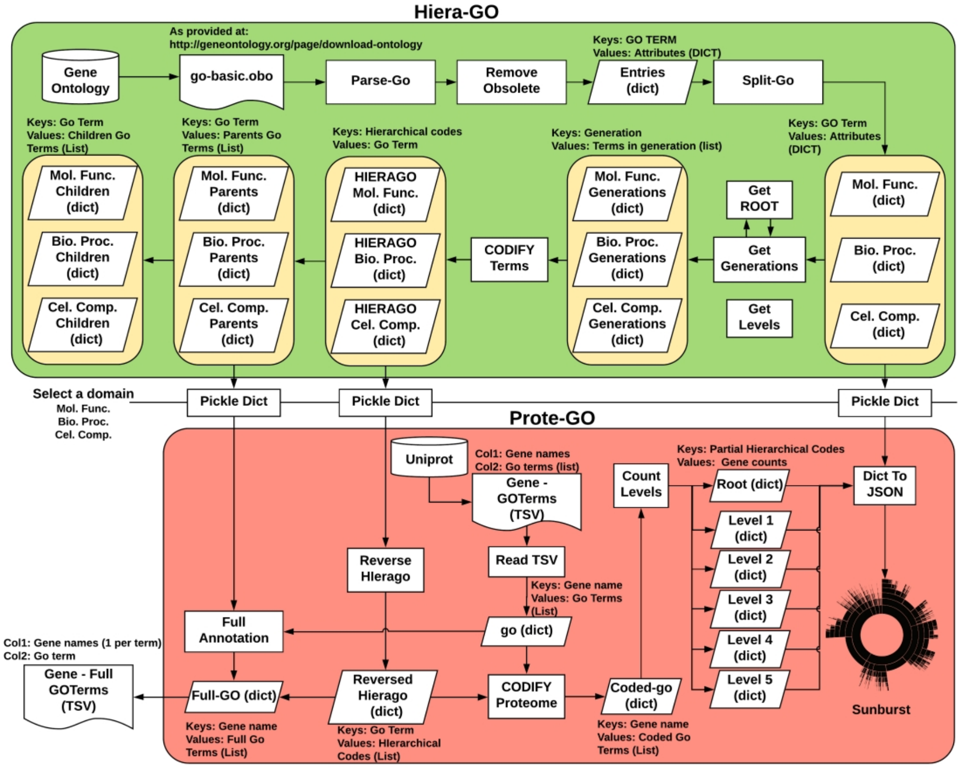


Figure 2 - Algorithm used to classify M. abscessus Gene Ontology terms as an interactive sunburst graph. The first set of functions was termed Hiera-GO and creates a set of codified GO-term dictionaries, one for each domain. The second set of functions was termed Prote-GO and uses the created dictionaries to produce the sunburst data and the complete GO annotation for each M. abscessus gene.

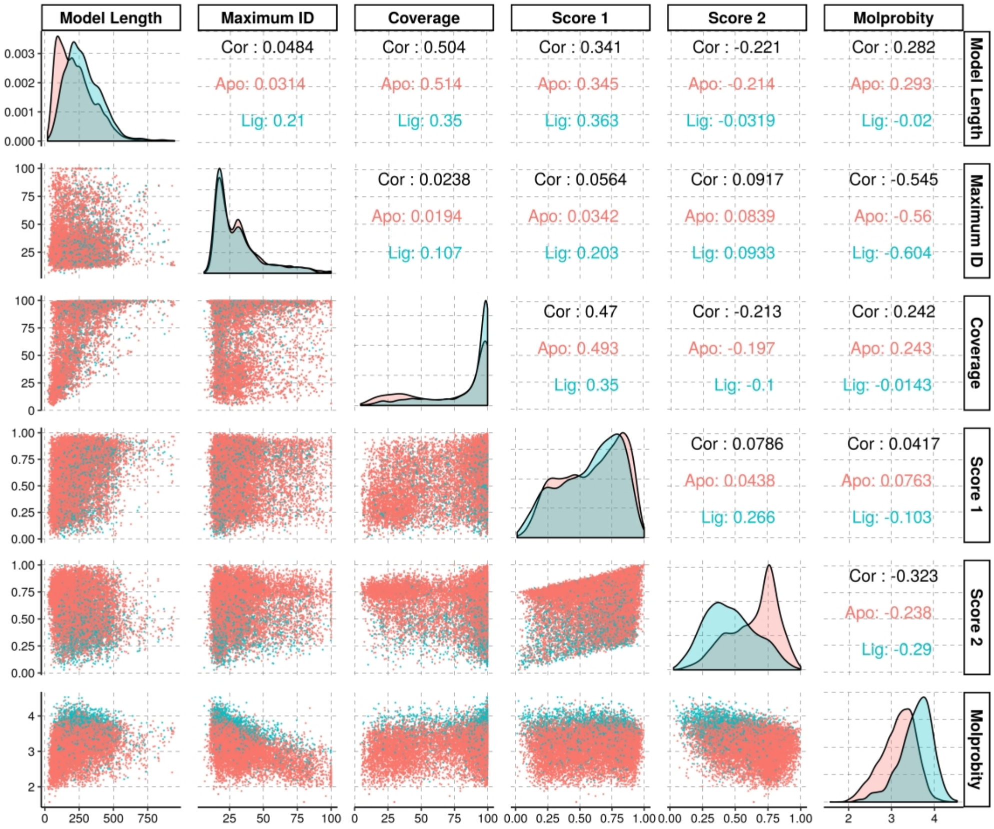


Figure 3 - Density and scatter plots of the properties calculated for each of the models. The colours represent the presence (cyan) or absence (red) of a ligand. Density plots are shown in the diagonal representing the distribution of each individual property. The bottom half contains the scatter plots and the top half of the graph presents the correlation coefficients.

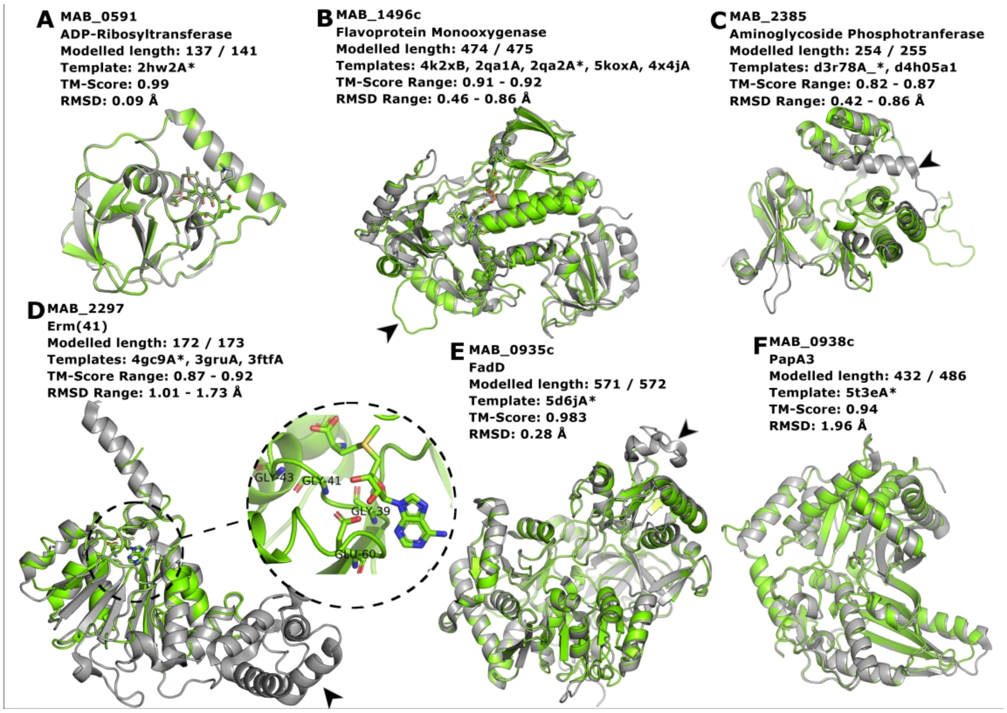
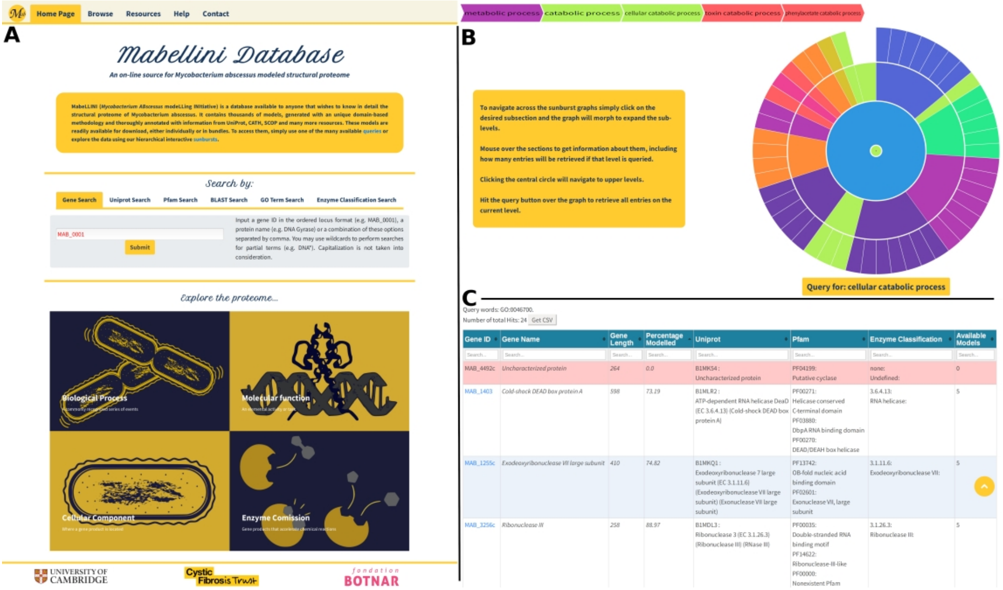

A MAB_0591
ADP-Ribosyltransferase
Modelled length: 137 / 141
Template: 2hw2A*
TM-Score: 0.99
RMSD: 0.09 Å
B MAB_1496c
Flavoprotein Monooxygenase
Modelled length: 474 / 475
Templates: 4k2xB, 2qa1A, 2qa2A*, 5koxA, 4x4jA
TM-Score Range: 0.91 - 0.92
RMSD Range: 0.46 - 0.86 Å
C MAB_2385
Aminoglycoside Phosphotranferase
Modelled length: 254 / 255
Templates: d3r78A_*, d4h05a1
TM-Score Range: 0.82 - 0.87
RMSD Range: 0.42 - 0.86 Å
D MAB_2297
Erm(41)
Modelled length: 172 / 173
Templates: 4gc9A*, 3gruA, 3ftfA
TM-Score Range: 0.87 - 0.92
RMSD Range: 1.01 - 1.73 Å
GLY-43
GLY-41
GLY-39
GLU-60
E MAB_0935c
FadD
Modelled length: 571 / 572
Template: 5d6jA*
TM-Score: 0.983
RMSD: 0.28 Å
F MAB_0938c
PapA3
Modelled length: 432 / 486
Template: 5t3eA*
TM-Score: 0.94
RMSD: 1.96 Å

Figure 4 - Interesting targets selected from the literature. (A-C) Antibiotic modifying proteins revised in (22). (D) Erythromycin ribosomal methylase (Erm(41)) with the SAM binding pocket and the GxGxG motif in detail. (E-F) FadD and PapA3; enzymes likely involved in the TPP-biosynthesis pathway as suggested in (32). In all cases, the models are represented in green and are superimposed to the most dissimilar template (i.e. the one presenting highest RMSD value), which is also indicated in each header with an asterisk. The black arrowheads point to different features described in the main text.



Mabellini Database

An on-line source for Mycobacterium abscessus modeled structural proteome

MabeLLINI (*Mycobacterium ABscessus* modeLLing INItiative) is a database available to anyone that wishes to know in detail the structural proteome of Mycobacterium abscessus. It contains thousands of models, generated with an unique domain-based methodology and thoroughly annotated with information from UniProt, CATH, SCOP and many more resources. These models are readily available for download, either individually or in bundles. To access them, simply use one of the many available queries or explore the data using our hierarchical interactive sunbursts.

Input a gene ID in the ordered locus format (e.g. MAB_0001), a protein name (e.g. DNA Gyrase) or a combination of these options separated by comma. You may use wildcards to perform searches for partial terms (e.g. DNA*). Capitalization is not taken into consideration.

To navigate across the sunburst graphs simply click on the desired subsection and the graph will morph to expand the sub-levels.

Mouse over the sections to get information about them, including how many entries will be retrieved if that level is queried.

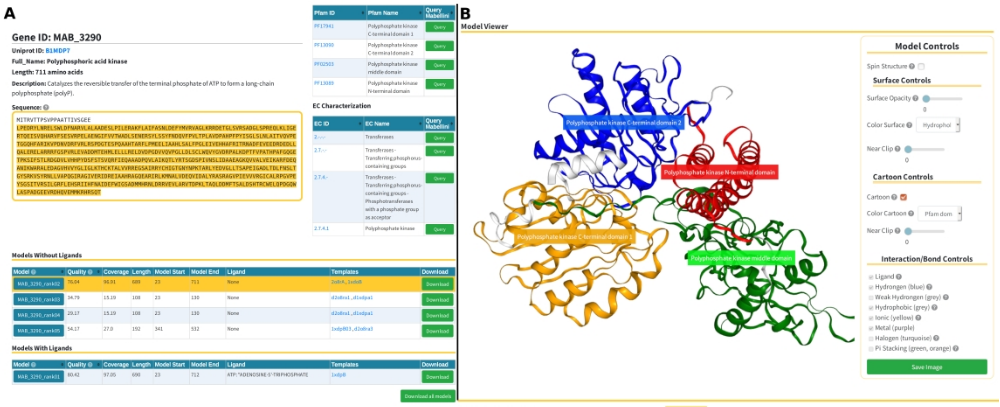
Clicking the central circle will navigate to upper levels.

Hit the query button over the graph to retrieve all entries on the current level.

Query words: GO:0046700.
Number of total Hits: 24

Gene ID	Gene Name	Gene Length	Percentage Modelled	Uniprot	Pfam	Enzyme Classification	Available Models
Search...	Search...	Search...	Search...	Search...	Search...	Search...	Search...
MAB_4492c	*Uncharacterized protein*	*264*	*0.0*	B1MK54 : Uncharacterized protein	PF04199: Putative cyclase	none: Undefined:	0
MAB_1403	*Cold-shock DEAD box protein A*	*598*	*73.19*	B1MLR2 : ATP-dependent RNA helicase DeaD (EC 3.6.4.13) (Cold-shock DEAD box protein A)	PF00271: Helicase conserved C-terminal domain PF03880: DbpA RNA binding domain PF00270: DEAD/DEAH box helicase	3.6.4.13: RNA helicase:	5
MAB_1255c	*Exodeoxyribonuclease VII large subunit*	*410*	*74.82*	B1MKQ1 : Exodeoxyribonuclease 7 large subunit (EC 3.1.11.6) (Exodeoxyribonuclease VII large subunit) (Exonuclease VII large subunit)	PF13742: OB-fold nucleic acid binding domain PF02601: Exonuclease VII, large subunit	3.1.11.6: Exodeoxyribonuclease VII:	5
MAB_3256c	*Ribonuclease III*	*258*	*88.97*	B1MDL3 : Ribonuclease 3 (EC 3.1.26.3) (Ribonuclease III) (RNase III)	PF00035: Double-stranded RNA binding motif PF14622: Ribonuclease-III-like PF00000: Nonexistent Pfam	3.1.26.3: Ribonuclease III:	5

Figure 5 - (A) Home Page, showing all query options and providing links to the sunburst navigation pages. (B) Sunburst representation of the Gene Ontology terms available for the M. abscessus proteome. The data were retrieved from UniProt and hierarchically classified using HieraGo. The graph is fully interactive and poses as an alternative querying system. Such classifications are available for the cellular component, molecular function and biological process ontology terms, as well as for Enzyme Commission numbers. (C) Searchable results table, showing information about the genes recovered by the initial query. Genes for which no models were produced are highlighted in red.


A
B
Polyphosphate kinase C-terminal domain 2
Polyphosphate kinase N-terminal domain
Polyphosphate kinase C-terminal domain 1
Polyphosphate kinase middle domain

Gene ID: MAB_3290

Uniprot ID: B1MDP7

Full_Name: Polyphosphoric acid kinase

Length: 711 amino acids

Description: Catalyzes the reversible transfer of the terminal phosphate of ATP to form a long-chain polyphosphate (polyP).

Sequence:

```
MITRVTTPSVPPAATTIVSGEE
LPEDRYLNRELSWLDFNARVLALAADESLPILERAKFLAIFASNLDEFYMVRVAGLKRRDETGLSVRSADGLSPREQLKLIGE
RTQEISVQHARVFSESVRPELAENGIFVVTWADLSENERSYLSSYFNDQVFPVLTPLAVDPAHPFPYISGLSLNLAITVQVPE
TGGQHFARIKVPDNVDRFVRLRSPDGTESPQAAHTARFLPMEELIAAHLSALFPGLEIVEHHAFRITRNADFEVEEDRDEDLL
QALERELARRRFGSPVRLEVADDMTEHMLELLLRELDVDPGDVVQVPGLLDLSCLWQVYGVDRPALKDPTFVPATHPAFGQGE
TPKSIFSTLRDGDVLVHHPYDSFSTSVQRFIEQAAADPQVLAIKQTLYRTSGDSPIVNSLIDAAEAGKQVVALVEIKARFDEQ
ANIKWARALEDAGVHVVYGLIGLKTHCKTALVVRREGSAIRRYCHIGTGNYNPKTARLYEDVGLLTSAPEIGADLTDLFNSLT
GYSRKVSYRNLLVAPQGIRAGIVERIDREIAAHRAGQEARIRLKMNALVDEQVIDALYRASRAGVPIEVVVRGICALRPGVPE
YSGSITVRSILGRFLEHSRIIHFNAIDEFWIGSADMMHRNLDRRVEVLARVTDPKLTAQLDDMFTSALDSHTRCWELQPDGQW
LASPADGEEVRDHQVEMMKRHRSQT
```

Pfam ID	Pfam Name	Query Mabellini
PF17941	Polyphosphate kinase C-terminal domain 1	Query
PF13090	Polyphosphate kinase C-terminal domain 2	Query
PF02503	Polyphosphate kinase middle domain	Query
PF13089	Polyphosphate kinase N-terminal domain	Query

EC Characterization

EC ID	EC Name	Query Mabellini
2.-.-.-	Transferases	Query
2.7.-.-	Transferases - Transferring phosphorus-containing groups	Query
2.7.4.-	Transferases - Transferring phosphorus-containing groups - Phosphotransferases with a phosphate group as acceptor	Query
2.7.4.1	Polyphosphate kinase	Query

Models Without Ligands

Model	Quality	Coverage	Length	Model Start	Model End	Ligand	Templates	Download
MAB_3290_rank02	76.04	96.91	689	23	711	None	2o8rA , 1xdoB	Download
MAB_3290_rank03	34.79	15.19	108	23	130	None	d2o8ra1 , d1xdpa1	Download
MAB_3290_rank04	29.17	15.19	108	23	130	None	d2o8ra1 , d1xdpa1	Download
MAB_3290_rank05	54.17	27.0	192	341	532	None	1xdpB03 , d2o8ra3	Download

Models With Ligands

Model	Quality	Coverage	Length	Model Start	Model End	Ligand	Templates	Download
MAB_3290_rank01	80.42	97.05	690	23	712	ATP:"ADENOSINE-5'-TRIPHOSPHATE	1xdpB	Download

Download all models

Model Viewer

Model Controls

Spin Structure ☐

Surface Controls

Surface Opacity 0

Color Surface Hydrophol

Near Clip 0

Cartoon Controls

Cartoon ☑

Color Cartoon Pfam dom

Near Clip 0

Interaction/Bond Controls

☑ Ligand
☑ Hydrongen (blue)
☐ Weak Hydrongen (grey)
☑ Hydrophobic (grey)
☑ Ionic (yellow)
☑ Metal (purple)
☐ Halogen (turquoise)
☐ Pi Stacking (green, orange)

Save Image

Figure 6 - Models page. (A) Information about the selected protein in the top section (UniProt ID, enzyme commission number, Pfam domain, protein sequence and description), followed by tables with the models' attributes for both liganded and non-liganded models. (B) In the bottom section, an NGL model viewer is displayed with a number of interactive options. The models are available for direct download, both individually or as a bundle.

Table 1 - Features of the best models built for proteins selected from the literature

Name	Prot. Len.	Model Len.	Cov. (%)	Max. Id. (%)	Score1	Score2	Molprob.	Relevance	Reference
MAB_0591	141	137	97.2	64.5	0.722	0.616	2.861	Rifamycin Resistance	(22)
MAB_0935c	572	571	99.8	39.3	0.936	0.571	3.159	TPP Biosyntetic Pathway	(32)
MAB_0938c	486	432	88.9	15.3	0.836	0.811	3.347		
MAB_1496c	475	474	99.8	32.4	0.853	0.442	3.817	Tetracycline Resistance	(22)
MAB_2297	173	172	99.4	23	0.845	0.48	3.861	Macrolide Resistance	(14), (22)
MAB_2385	255	254	99.6	32.2	0.963	0.837	2.87	Aminoglycoside Resistance	(22)
MAB_2875	289	245	84.8	99.6	0.504	0.71	2.258	β-Lactam Resistance	(22)
MAB_4100c	76	75	98.7	82.7	0.451	0.415	2.809	GPL Biosynthetic Pathway	(64)
MAB_4103c	267	225	84.3	20.2	0.6	0.381	4.024		
MAB_4105c	262	236	90.1	23	0.433	0.758	3.587		
MAB_4108c	268	262	97.8	47.2	0.683	0.312	3.041		
MAB_4109c	230	229	99.6	28.3	0.888	0.643	3.567		
MAB_4111c	353	348	98.6	31.2	0.926	0.923	3.345		
MAB_4112c	440	418	95	19.3	0.96	0.684	3.44		
MAB_4113	288	287	99.7	60.1	0.887	0.833	2.651		
MAB_4115c	987	331	33.5	16.3	0.82	0.405	3.421		
MAB_4116c	959	939	97.9	13.5	0.964	0.567	3.548		
MAB_4117c	138	89	64.5	50	0.778	0.865	2.954		

Mabellini: a genome-wide database for understanding the structural proteome and evaluating prospective antimicrobial targets of the emerging pathogen *Mycobacterium abscessus*

Supplementary Material

Marcin J. Skwark[1¶], Pedro H. M. Torres[1¶], Liviu Copoiu[1¶], Bridget Bannerman[1], R. Andres Floto[2,3] and Tom L. Blundell[1*]

[1] *Department of Biochemistry, University of Cambridge, Cambridge, CB2 1GA, UK*

[2] *Molecular Immunity Unit, Department of Medicine University of Cambridge, MRC-Laboratory of Molecular Biology, Cambridge, CB2 0QH, UK*

[3] *Cambridge Centre for Lung Infection, Royal Papworth Hospital, Cambridge, CB23 3RE, UK*

*Corresponding Author
E-mail: tlb20@cam.ac.uk (TLB)

¶These authors have contributed equally to this manuscript

AEROSPACI scores supplementation

Both VIVACE and TOCCATA are fundamentally based on a SCOP+CATH combination of domain annotations and metadata. One of these are AEROSPACI scores present in SCOP. Nevertheless, a significant fraction of structures present in PDB at the time of preparation of this work were devoid of AEROSPACI scores. To circumvent this, we have devised a simple, linear model:

$$A_s(\boldsymbol{x}) = \sum_{i=0}^{4} \beta_i \, x_i$$

where x designates observed values, β designates the fit parameters (see below) and indices correspond respectively to: resolution, reciprocal of resolution, *rvalue*, reciprocal of *rvalue* and ratio of resolution to *rvalue*. We have used the values present in the SCOP "aerospaci-all-data-2018-05-17.txt" file and applied a grid search algorithm to find the best set of hyperparameters for the Lasso algorithm using the scikit-learn python module. For hyperparameters of $\alpha = 1e\text{-}5$, $max\_iter = 5000$, we found that parameters $\beta = [-0.04354, 0.91651, -1.03199, 0.00031, -0.00048]$ yielded the best goodness of fit to observed values ($r^2 = 0.956$). We explored other regression methods (Ridge Regression and simple linear regression with stochastic gradient descent. The hyperparameter search was performed in a randomized, five-fold cross-validated regime.

The fit parameters have been applied to fill in the AEROSPACI values missing for certain PDBs.

API Usage

The currently implemented API includes queries by (i) Identifiers (Pfam, Ordered Locus, UniProt, Enzyme Comission, Gene Ontology), (ii) text, (iii) ligand ID, (iv) single model retrieval and (v) best models. The queries return JSON files that can be easily parsed programmatically containing information about the genes and models. The usage is described below.

Search by identifier:

Each of the identifiers is case insensitive. A query takes a single identifier and returns the information about gene (or genes) and links to the model JSON files.

Usage
http://mabellinidb.science/api/gene/[ID] - for ordered locus names
http://mabellinidb.science/api/oln/[ID] - alias for the former
http://mabellinidb.science/api/uniprot/[ID] - for UniProt IDs
http://mabellinidb.science/api/pfam/[ID] - for Pfam IDs
http://mabellinidb.science/api/EC/[ID] - for EC IDs
http://mabellinidb.science/api/GO/[ID] - for GO terms
Examples:
http://mabellinidb.science/api/gene/MAB_1234

Free text search:

Case sensitive search through the free-form fields of the database, that is names of genes and descriptions. The text is not tokenized, thus out-of-order search does not match.

Usage
http://mabellinidb.science/api/text/[text]
Example
http://mabellinidb.science/api/text/cell%20wall

Single model retrieval:

Given the internal model ID (returned by the search function above), return the URL of model's PDB and mmCIF files, as well as model descriptors in terms of quality, templates and ligand status.

Usage
http://mabellinidb.science/api/model/[ID]
Example
http://mabellinidb.science/api/model/MAB_1668__rank01__apo__1-341__Q_0.836_0.378.pdb

The function also recognizes queries in a form [ordered locus name]_[model rank]. The example below would retrieve the first ranked model for MAB_1234.

Example:
http://mabellinidb.science/api/model/MAB_1234__1

Retrieval of model IDs by ligand:

Given a three-letter ligand code (as per PDB small molecule dictionary), returns URLs for all the JSON records for models containing the ligand.

Usage
http://mabellinidb.science/api/ligand/[ID]
Example
http://mabellinidb.science/api/ligand/COA

Best models:

A convenience, parameter-free function, returning URLs of all the JSON records for the first ranked models.

Usage
http://mabellinidb.science/api/bestModels

Supplementary Figures

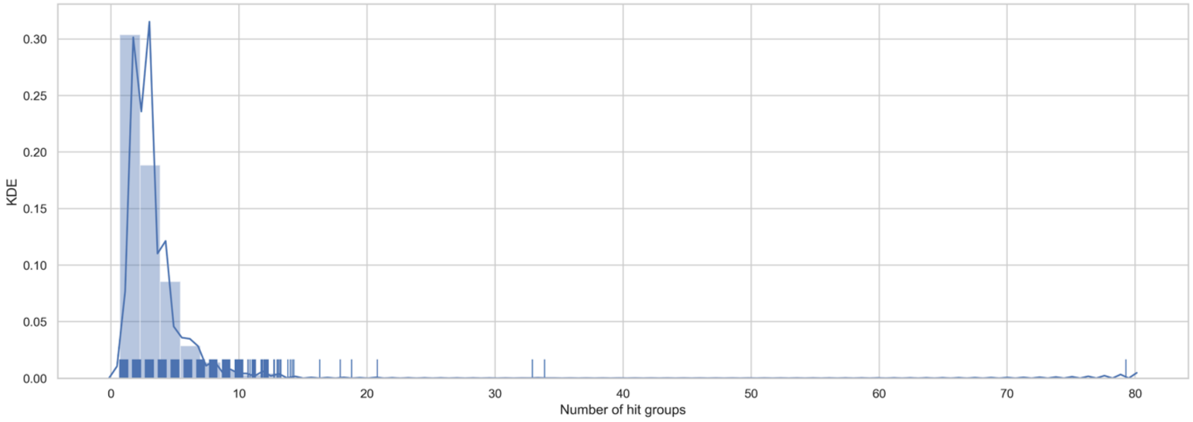


Figure S1 - The distribution of the number of TOCCATA hit groups among the Mabellini targets. For most of the targets, few (<5) hit groups have been identified, indicating a high confidence of the identified hits. Each of the ticks in the rug plot on the X-axis corresponds to the Mabellini target protein and the number of hit groups identified therein. Y-axis corresponds to the kernel density estimate. The evident outlier in the right-hand side of the plot is MAB_4691c (non-ribosomal peptide synthetase PstA) with 79 hit groups.

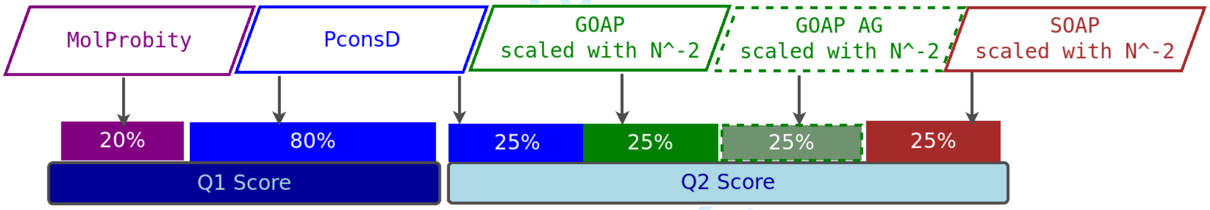


Figure S2 - Composition of Q1 and Q2 scores. Q1 is a linear combination of 80% structural consensus between models (computed by PconsD) and 20% stereochemical quality (percentage score). Q2 is a linear combination of GOAP-AG54, GOAP-score and SOAP55 potentials (scaled by the reciprocal of the square of number of residues in the model) and PconsD score. All four metrics composing the Q2 score have equal weights.

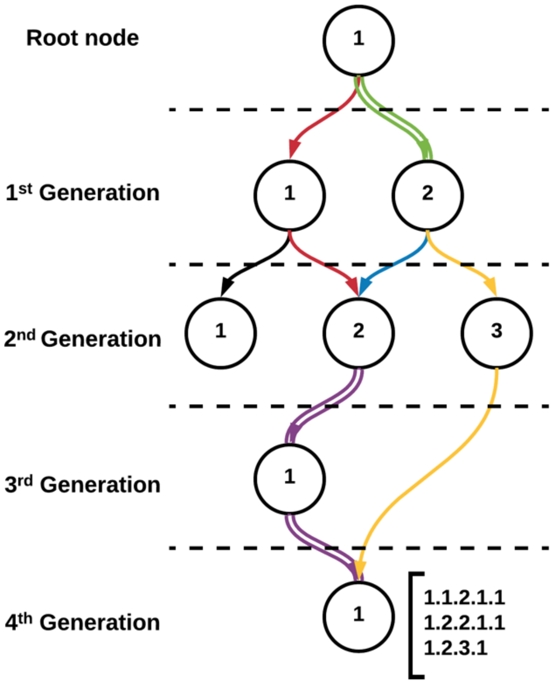


Figure S3 - Schematics of the hierarchical codes assigned to each GO term. Every node is assigned codes that represent every possible path from the root node. Also, the nodes in each layer (or generation) are numbered sequentially. These coded terms are then used to generate the hierarchical representation used in the interactive sunbursts (c.f. Figure 2).

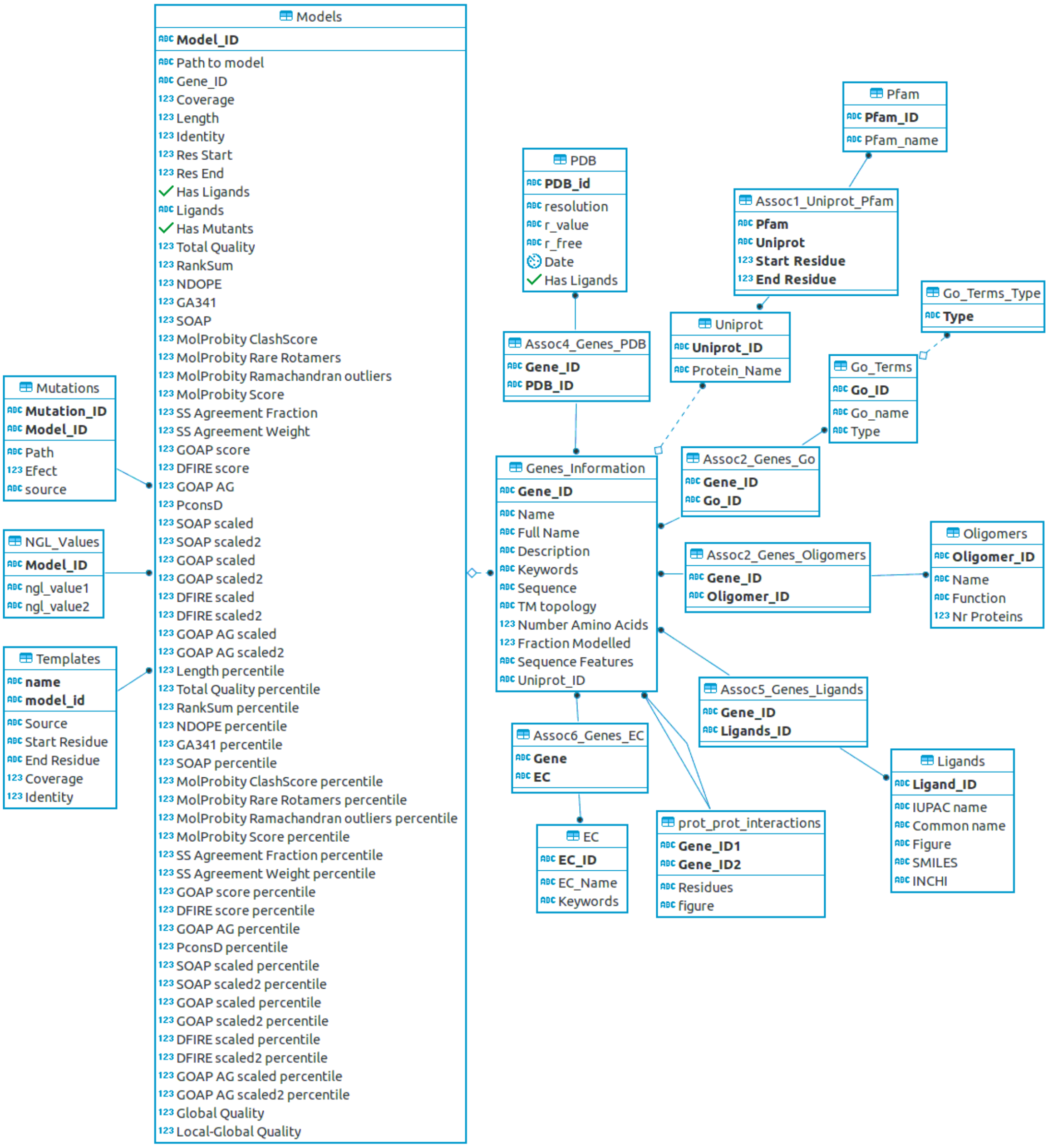


Figure S4 - Database Schema as implemented using SQLAlchemy. The *Genes Information* table consists the central hub connecting the bulk of the data. The models table comprises all the available information for the created models (quality, length, span, coverage, identity, etc.). A total of 5 association tables are in place to solve the many-to-many relationships (Uniprot-Pfam, Gene-GO Term, Gene-PDB, Gene-Ligands and Gene-EC). The NGL table contains information to be displayed in the NGL Viewer implemented on the website, such as the coordinates for the centres of mass of the Pfam domains. The database is already suited to receive the idealized updates, such as the effects of mutations and the oligomeric structures.