MoonDB 2.0: an updated database of extreme multifunctional and moonlighting proteins
Diogo Ribeiro, Galadriel Briere, Benoit Bely, Lionel Spinelli, Christine Brun

To cite this version:
Diogo Ribeiro, Galadriel Briere, Benoit Bely, Lionel Spinelli, Christine Brun. MoonDB 2.0: an updated database of extreme multifunctional and moonlighting proteins. Nucleic Acids Research, 2019, 10.1093/nar/gky1039. hal-01907669

HAL Id: hal-01907669
https://amu.hal.science/hal-01907669
Submitted on 29 Oct 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License
MoonDB 2.0: an updated database of extreme multifunctional and moonlighting proteins
Diogo Ribeiro, Galadriel Briere, Benoit Bely, Lionel Spinelli, Christine Brun

To cite this version:
Diogo Ribeiro, Galadriel Briere, Benoit Bely, Lionel Spinelli, Christine Brun. MoonDB 2.0: an updated database of extreme multifunctional and moonlighting proteins. Nucleic Acids Research, Oxford University Press, 2018, <10.1093/nar/gky1039>. <hal-01907669>
MoonDB 2.0: an updated database of extreme multifunctional and moonlighting proteins

Diogo M. Ribeiro¹, Galadriel Briere¹, Benoit Bely², Lionel Spinelli¹ and Christine Brun ¹,³, *

¹Aix-Marseille Univ, INSERM, TAGC, UMR_S1090, Marseille, France, ²The European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus CB10 1SD, UK and ³CNRS, Marseille, France

ABSTRACT
MoonDB 2.0 (http://moondb.hu.univ-amu.fr/) is a database of predicted and manually curated extreme multifunctional (EMF) and moonlighting proteins, i.e. proteins that perform multiple unrelated functions. We have previously shown that such proteins can be predicted through the analysis of their molecular interaction subnetworks, their functional annotations and their association to distinct groups of proteins that are involved in unrelated functions. In MoonDB 2.0, we updated the set of human EMF proteins (238 proteins), using the latest functional annotations and protein–protein interaction networks. Furthermore, for the first time, we applied our method to four additional model organisms - mouse, fly, worm and yeast - and identified 54 novel EMF proteins in these species. In addition to novel predictions, this update contains 63 human and yeast proteins that were manually curated from literature, including descriptions of moonlighting functions and associated references. Importantly, MoonDB’s interface was fully redesigned and improved, and its entries are now cross-referenced in the UniProt Knowledgebase (UniProtKB). MoonDB will be updated once a year with the novel EMF candidates calculated from the latest available protein interactions and functional annotations.

INTRODUCTION
Moonlighting, multitask and extreme multifunctional proteins are proteins that perform multiple unrelated biological functions, regardless of their domain organisation and their evolutionary history (1–3). A canonical example of a moonlighting protein is the human aconitase, an enzyme of the tricarboxylic acid cycle (TCA cycle) that also functions as a translational regulator, upon a conformational change (4). Extreme multifunctional and moonlighting proteins are present throughout the evolutionary tree, and their unrelated functions may be performed in different tissues or cellular locations, sometimes associated (either as a cause or a consequence) to a change in their interaction partners, conformation or oligomeric states (5). These proteins are often in the intersection—and may coordinate—several pathways or responses to different stimuli (6). Despite their importance, the moonlighting functions of proteins have usually been identified by serendipity, since clear procedures to identify secondary functions have not been proposed. As a consequence, the prevalence of moonlighting proteins in proteomes was unknown. This prompted us to provide in 2015, MoonGO, a computational pipeline to identify extreme multifunctional (EMF) proteins on a large scale (2). EMF proteins were identified by exploiting the topology of protein–protein interaction networks and protein GO term annotations, without any a priori knowledge of moonlighting. The first version of MoonDB (2) contained the EMF proteins predicted by MoonGO, complemented with a careful manual curation of literature of moonlighting or EMF proteins. Here, we present MoonDB 2.0, an update that, besides improving predictions and manual curation for human, also includes predicted and curated entries for four other model organisms - mouse, fly, worm and yeast. Our main focus is to provide users with an extensive set of predicted and curated EMF and moonlighting proteins, describing their functions comprehensively.

MATERIALS AND METHODS
Prediction of extreme multifunctional proteins
The method used to predict extreme multifunctional (EMF) proteins was first described in Chapple et al. (2). Briefly, we perform a large-scale search for EMF proteins by (i) identifying functionally-dissimilar pairs of Biological Process Gene Ontology (GO) terms with PrOnto (7) that uses two metrics of GO functional dissimilarity; (ii) clustering the protein interactome into overlapping clusters of proteins using the OCG algorithm (8); (iii) annotating each cluster with functions (Biological Pro-
cess GO terms) based on the annotations of its constituent proteins; (iv) identifying proteins that belong to at least two clusters and are annotated to dissimilar functions (after having inherited the annotations of their clusters in addition to their own), hereby labeled as EMF proteins. We used protein–protein interaction data gathered on December 2017 from the PSICQUIC webservice (9), processed as described in Chapple et al. 2015 (2). We only include experimentally identified binary protein–protein interactions, by considering only interactions from certain experimental methods (Supplementary Table S2). GO term annotations and ontologies were collected from the Gene Ontology Consortium (10) on December 2017.

Criteria for manual curation

We provide a list of bona fide moonlighting and extreme multifunctional proteins manually curated from literature over the years. Each entry was confirmed independently by at least two members of our team. Specifically, we confirm that the several functions are indeed distinct to each other and not a by-product of the same function under different circumstances (e.g., regulation of two distinct pathways through the same mechanism, such as phosphorylation). In each case, publications describing the different functions of a protein are provided. When available, the conditions that may be related to the change in function are also described (e.g., cellular localisation, oligomerization).

Database architecture and web interface

MoonDB 2.0 has been developed using the SQLAlchemy (v1.2.0) Python (v2.7.6) library for data storage. The web interface is mainly written in PHP (v7.1.14) and Jquery (v3.2.1) and is powered by the Drupal (v8.4) Content Management System (CMS). The database was deployed with Docker (v17.09.0-ce) to ensure stability. We gathered information on protein domains, publications and diseases fromUniProtKB (11) in January 2018.

DATABASE CONTENT AND WEB INTERFACE

A new dataset of extreme multifunctional proteins

In MoonDB 2.0 we have predicted 292 extreme multifunctional (EMF) proteins in human, and—for the first time—also in mouse, fly, worm and yeast species (Supplementary Table S1). These have been identified de novo using the latest protein–protein interaction networks and GO term annotations. The power to predict EMF proteins is dependent on the underlying coverage and quality of the protein interactomes and GO term annotations used (Supplementary Table S1). The new human EMF protein dataset and the one in the previous MoonDB version overlap significantly (Supplementary Figure S1). Interestingly, the analysis of the EMF signature on the new set of EMF proteins (as performed in Chapple et al. (2)), shows that the new set of EMF proteins produces a similar signature in terms of network properties, tissue expression, as well as domain, structural disorder and Eukaryotic linear motif (ELM) presence, among other features (Supplementary Figure S2). Notably, as observed in other studies (2,12,13), moonlighting and EMF proteins are often associated to disease, a feature also observed for the set of human EMF proteins when considering disease-associations from OMIM (P = 2.2 × 10⁻²⁸, odds ratio = 2.10; one-tailed Fishers’s Exact test).

Besides EMF predictions, in this update we also manually curated 15 yeast moonlighting proteins, describing their unrelated functions, specifying which conditions may influence moonlighting (e.g., cellular localisation) and referencing relevant publications. Similarly, we added functional descriptions for 47 human moonlighting proteins. All these proteins constitute the ‘Reference Set’.

A new user-friendly interface and additional content

To provide our visitors with a clear, fast and easy-to-use database, we completely redesigned MoonDB’s web interface and added new functionalities. It is now possible to search a MoonDB entry by gene name, UniProtKB identifier (ID) or UniProtKB accession (AC). Moreover, the ‘Browse’ page (Figure 1), which displays all protein entries in MoonDB, can now be filtered through any column, thus allowing searches by species, full name of the protein and its presence in the ‘Reference Set’. These filters can also be used in combination with each other to make more advanced queries.

Importantly, MoonDB specifies which pairs of dissimilar (i.e., unrelated) functions led us to propose each predicted EMF and curated protein as moonlighting/extreme-multifunctional (Figure 2, under ‘MoonDB Dissimilar Functions’). Furthermore, we provide the set of GO terms associated to the protein (Figure 2, under ‘MoonDB Network Modules’), determined by its participation in network clusters with annotated functions (guilt-by-association principle), and the GO terms directly annotating the protein. This information is pertinent in the context of multifunctionality, since EMF proteins associate with several groups of proteins to perform alternative functions. In addition, since the ability to perform unrelated functions may be correlated with the presence of a protein in unrelated subcellular locations, in MoonDB 2.0 we identified pairs of unrelated cellular component GO terms associated to each protein with PrOnto (7) (Figure 2, under ‘Protein GO Annotations’). Lastly, to fully describe moonlighting and extreme multifunctional proteins, we further cross-link functional data with other orthogonal information such as the protein association to disease, protein domains and publications associated to the protein.

DISCUSSION AND CONCLUSION

The MoonDB 2.0 database is accessible at http://moondb.hb.univ-amu.fr/ and now contains data for human, mouse, fly, worm and yeast. MoonDB 2.0 stands out compared to the two other current databases of moonlighting proteins MoonProt (14) and Multitask-II (12) because MoonDB 2.0 combines curated and predicted proteins. We consider our dataset to be more comprehensive as well as highly complementary to other available databases. Whereas other databases are dependent on the available literature, and thus limited to providing information which is already known,
Figure 1. MoonDB 2.0 browse page. The browse page displays the entries of all MoonDB 2.0 proteins and can be searched interactively. The 'MoonDB ID' can be clicked to access each individual MoonDB 2.0 protein entry.

our dataset of predictions goes beyond current propositions of moonlighting and provides novel candidates. EMF prediction is large-scale and detection does not require *a priori* knowledge besides protein interactions and GO term annotations.

Importantly, as protein interactomes and GO term annotations of model organisms will continue to grow towards completion in the following years, MoonDB will be updated every year with EMF predictions made from the latest interactomes and GO term annotations. Both power and reliability will progressively increase with future releases. This will be particularly important in the cases of mouse, which possesses high-quality GO term annotations (average of ~19 GO terms per protein), but an incomplete protein interactome (<15% of the proteome covered), as well as in fly, whose interactome is better covered (>40% proteome), but GO term annotations are available for less than half of the interactome. Consequently, only few EMF proteins in mouse and fly were detected with our method. However, 5 out of 14 mouse EMF proteins are orthologs of human EMF proteins, suggesting that even when data is limited, the EMF proteins predicted are reliable. Indeed, the ability for genes to be multifunctional is conserved across orthologs of different organisms (16). Notably, orthologs were also found between human and worm (UBE2I/ubc-9 and SUMO1; 2 out of 6 MoonDB 2.0 worm entries) and even between the distant human and yeast species (SKP1 gene), although our method does not use ortholog relationships for EMF prediction. Together, these findings further underline the quality of our predictions and designates MoonDB 2.0 as a valuable data repository for one interested in studying the extreme multifunctionality and moonlighting of proteins, possibly across species.

We believe that MoonDB is of interest not only to bioinformaticians working on multifunctionality, but also to any biologist who may profit from knowing whether their protein of study is likely to perform unexpected functions aside from the ones generally known. Due to the high frequency of EMF proteins involved in multiple diseases, often in comorbidity (13), the extensive functional information provided in MoonDB 2.0 is of interest to help designing therapies that are aware of the several functions of the protein. Importantly, MoonDB 2.0 is now cross-referenced in the UniProt Knowledgebase (UniProtKB) (11). We consider that this greatly magnifies the exposure of our database to the general scientific community, as UniProtKB is the reference database for protein-related data and widely used by biologists, biochemists, bioinformaticians and others.
Figure 2. Example of a MoonDB’s protein entry. Protein entries provide extensive functional information such as the dissimilar function annotations and GO term annotations from network modules, as well as publications, diseases and domains associated with the protein.
DATA AVAILABILITY
MoonDB 2.0 is freely available at http://moondb.hb.univ-amu.fr/. Files containing EMF protein lists for each species, as well as the protein–protein interaction networks used in this study are freely available for download in MoonDB 2.0, and can be used in accordance with the GNU Public License and the license of primary data sources.

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS
We would like to thank Zacharie Menetrier for creating the MoonDB 2.0 logo, Benoit Ballester for tips regarding biological databases and Andreas Zanzoni for critically reading the manuscript and testing the database.

FUNDING

Conflict of interest statement. None declared.

REFERENCES