Journal of Biomedical Informatics 91 (2019) 103121

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Computational prediction of the bioactivity potential of proteomes based on expert knowledge

Aitor Blanco-Míguez[a,b,c], Guillermo Blanco[a,b,c], Alberto Gutierrez-Jácome[a], Florentino Fdez-Riverola[a,b,d], Borja Sánchez[c], Anália Lourenço[a,b,d,e,*]

[a] ESEI: Escuela Superior de Ingeniería Informática, University of Vigo, Edificio Politécnico, Campus Universitario As Lagoas, s/n, 32004 Ourense, Spain
[b] CINBIO - Centro de Investigaciones Biomédicas, University of Vigo, Campus Universitario Lagoas-Marcosende, 36310 Vigo, Spain
[c] Department of Microbiology and Biochemistry of Dairy Products, Instituto de Productos Lácteos de Asturias (IPLA), Consejo Superior de Investigaciones Científicas (CSIC), Paseo Río Linares, S/N, 33300 Villaviciosa, Asturias, Spain
[d] SING Research Group, Galicia Sur Health Research Institute (IIS Galicia Sur), SERGAS-UVIGO, Hospital Álvaro Cunqueiro, 36312 Vigo, Spain
[e] CEB - Centre of Biological Engineering, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal

ARTICLE INFO

Keywords:
Proteomes
Metaproteomes
Functionally relevant proteins
Bioactivity prediction
Translational application

ABSTRACT

Advances in the field of genome sequencing have enabled a comprehensive analysis and annotation of the dynamics of the protein inventory of cells. This has been proven particularly rewarding for microbial cells, for which the majority of proteins are already accessible to analysis through automatic metagenome annotation. The large-scale *in silico* screening of proteomes and metaproteomes is key to uncover bioactivities of translational, clinical and biotechnological interest, and to help assign functions to certain proteins, such as those predicted as hypothetical. This work introduces a new method for the prediction of the bioactivity potential of proteomes/metaproteomes, supporting the discovery of functionally relevant proteins based on prior knowledge. This methodology complements functional annotation enrichment methods by allowing the assignment of functions to proteins annotated as hypothetical/putative/uncharacterised, as well as and enabling the detection of specific bioactivities and the recovery of proteins from defined taxa.

This work shows how the new method can be applied to screen proteome and metaproteome sets to obtain predictions of clinical or biotechnological interest based on reference datasets. Notably, with this methodology, the large information files obtained after DNA sequencing or protein identification experiments can be associated for translational purposes that, in cases such as antibiotic-resistance pathogens or foodborne diseases, may represent changes in how these important and global health burdens are approached in the clinical practice.

Finally, the Sequence-based Expert-driven pRoteome bioactivity Prediction EnvironmENT, a public Web service implemented in Scala functional programming style, is introduced as means to ensure broad access to the method as well as to discuss main implementation issues, such as modularity, extensibility and interoperability.

1. Introduction

The functional annotation of the lists of genes and proteins derived from high-throughput experiments is a key, yet challenging Bioinformatics task [1]. The traditional strategy for this task is the functional enrichment analysis. Typically, enrichment approaches use the gene/protein annotations provided by consolidated biomedical ontologies or knowledge bases, such as Gene Ontology [2], UniProt [3] or Reactome [4], to infer which annotations are under- or over-represented in the list of genes or proteins under study. The assumption is that such enriched terms describe important underlying biological processes or behaviours [5].

Early in 2002 and 2003, several independent studies addressed the functional enrichment analysis of large lists of genes [6,7]. Onto-Express [8], MAPPFinder [9], GoMiner [10], DAVID [11], EASE [12], GeneMerge [13] and FuncAssociate [14] are among the first batch of successful tools. Since then, the field of functional enrichment analysis has been quite productive, resulting in a growing number of publicly available tools, typically categorised into three major classes, i.e. singular enrichment analysis (SEA), gene set enrichment analysis (GSEA), and modular enrichment analysis (MEA) [1].

This work introduces a complimentary method that specifically

* Corresponding author at: ESEI: Escuela Superior de Ingeniería Informática, University of Vigo, Edificio Politécnico, Campus Universitario As Lagoas, s/n, 32004 Ourense, Spain.
E-mail address: analia@uvigo.es (A. Lourenço).

https://doi.org/10.1016/j.jbi.2019.103121

Available online 07 February 2019

tackles the annotation of hypothetical, putative and uncharacterised proteins, which none of the conventional functional enrichment methods does, while expediting the screening of proteomes or metaproteomes for bioactivities of clinical and biotechnological interest. Traditional functional enrichment methods identify terms or modules significantly enriched in a set of genes or proteins. Instead, the proposed method looks for user-selected bioactivities through sequence similarity analysis. So, the new method is flexible to the interests of the user, i.e. triggering such screening in terms of sequences, bioactivities or taxa of interest, and information availability, i.e. the user may select the sequence references from molecular databases, literature or even in-house wet lab experiments.

Previous work has already shown the usefulness of a similar approach for the screening of immunomodulatory and proliferative peptides encrypted in the human gut microbiota [15]. Further noteworthy examples of the translational application of the method are antibiotic resistance and foodborne diseases. Since many of the global health burdens correspond to this kind of microbial diseases in which the aetiological agent can be identified by, at least, one specific protein, e.g. diarrheal diseases produced by cholera or Shiga-toxin producing *E. coli*, invasive enteric diseases produced by *Brucella* or *Listeria* spp., or even intoxications produced by *Clostridium botulinum* or *Bacillus* sp.[16], a method of bioactivity prediction based on expert knowledge seems to be a proper approach.

Therefore, the rest of this work is devoted to the description of new method as well as to the illustration of its rationale. The cases of study are two metaproteome screenings of clinical interest, i.e. the presence of taxon-specific prophages and the methicillin resistance protein MecA, and two well-documented proteome cases relevant to human gut microbiota research, i.e. the activity of glycoside hydrolases and tetracycline resistance protein Tet(W). The implementation of the method is also discussed. This methodology can be freely accessed at the Sequence-based Expert-driven pRoteome bioactivity Prediction EnvironmENT site (http://sing-group.org/serpent).

2. Materials and methods

2.1. Activity-driven functional enrichment methodology

The proposed method is specifically designed to enable the user to indicate the curated set of protein sequences to be used as reference dataset. That is, from the beginning, the knowledge discovery process targets a specific set of bioactivities of interest (Fig. 1). These sequences may be compiled from a variety of sources, including molecular databases, literature or wet lab experiments, and they may contain various functional annotations, derived from a biomedical ontology or based on the user's expertise. Likewise, the set of proteomes or metaproteomes to be analysed is specified according to the interests of analysis (i.e. a broader or more focused screening).

Bioactivity potential prediction is achieved based on how similar the proteomes or metaproteomes under analysis are to the reference sequences. Such analysis is based on the alignment distance matrix produced by a global progressive alignment process [17]. Each proteome is analysed individually by computing the distance between each of its proteins and each of the sequences in the reference dataset. The similarity threshold enables the identification of the proteomes holding higher bioactivity potential, and the likelihood of the bioactivities (if the reference set specifies more than one) is described as means to facilitate further empirical analysis. Those protein sequences that meet the similarity threshold specified by the user are collected and ordered by the best score obtained. Furthermore, statistical estimations of the predictions are calculated in order to control both the family-wise error rate (FWER) and the false discovery rate (FDR).

The level of confidence on the protein sequences under analysis depends on the specific interests of the user and the available information. As such, the results of the analysis should be interpreted

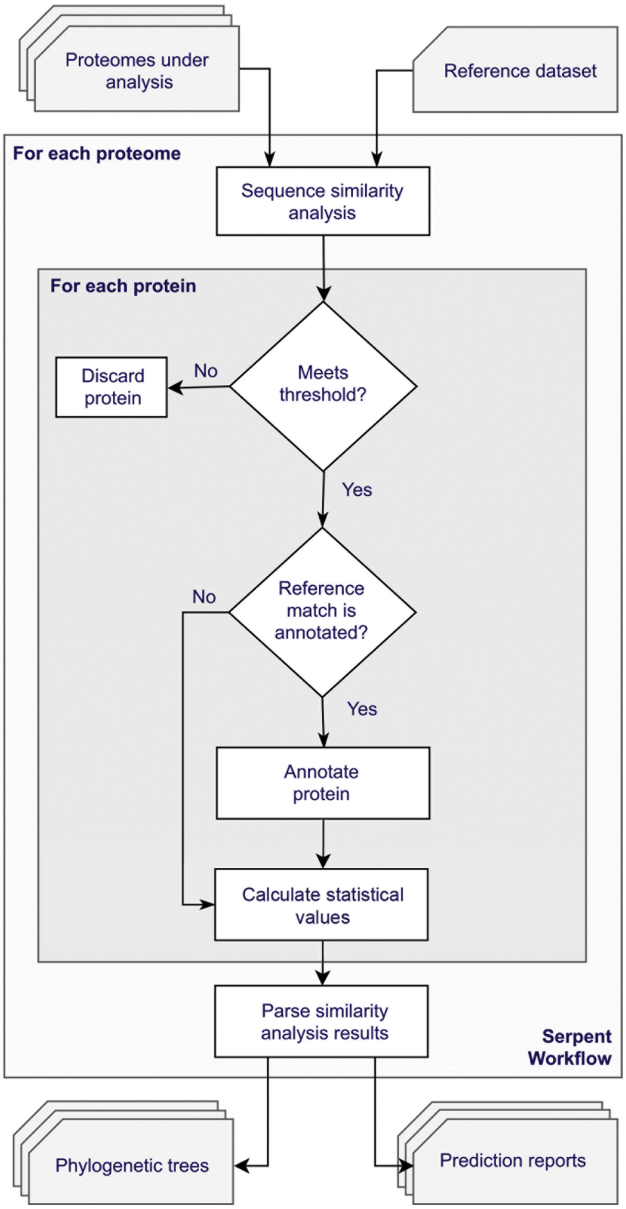


Fig. 1. Schematic of the workflow of the activity-driven bioactivity potential prediction methodology.

taking into account both the percentage of similarity between the sequences, and the statistical estimation of the predictions. To facilitate the visual exploration of results, the relationships observed between the reference sequences and the target sequences are represented in a phylogenetic tree.

2.2. Web service

Different software packages and programming languages can be used for implementing the proposed method. Here, attention is given to the Sequence-based Expert-driven pRoteome bioactivity Prediction EnvironmENT (Serpent), which provides an implementation in Scala programming language [18]. The rationale behind the Serpent design was to pursue a purely functional programming style while achieving the best possible efficiency. The most time-consuming parts of the workflow were developed so that it is possible to exploit the parallel capabilities of the current hardware systems, and thus reduce the time spent in each analysis.

The architecture is completely modular to facilitate the extension of

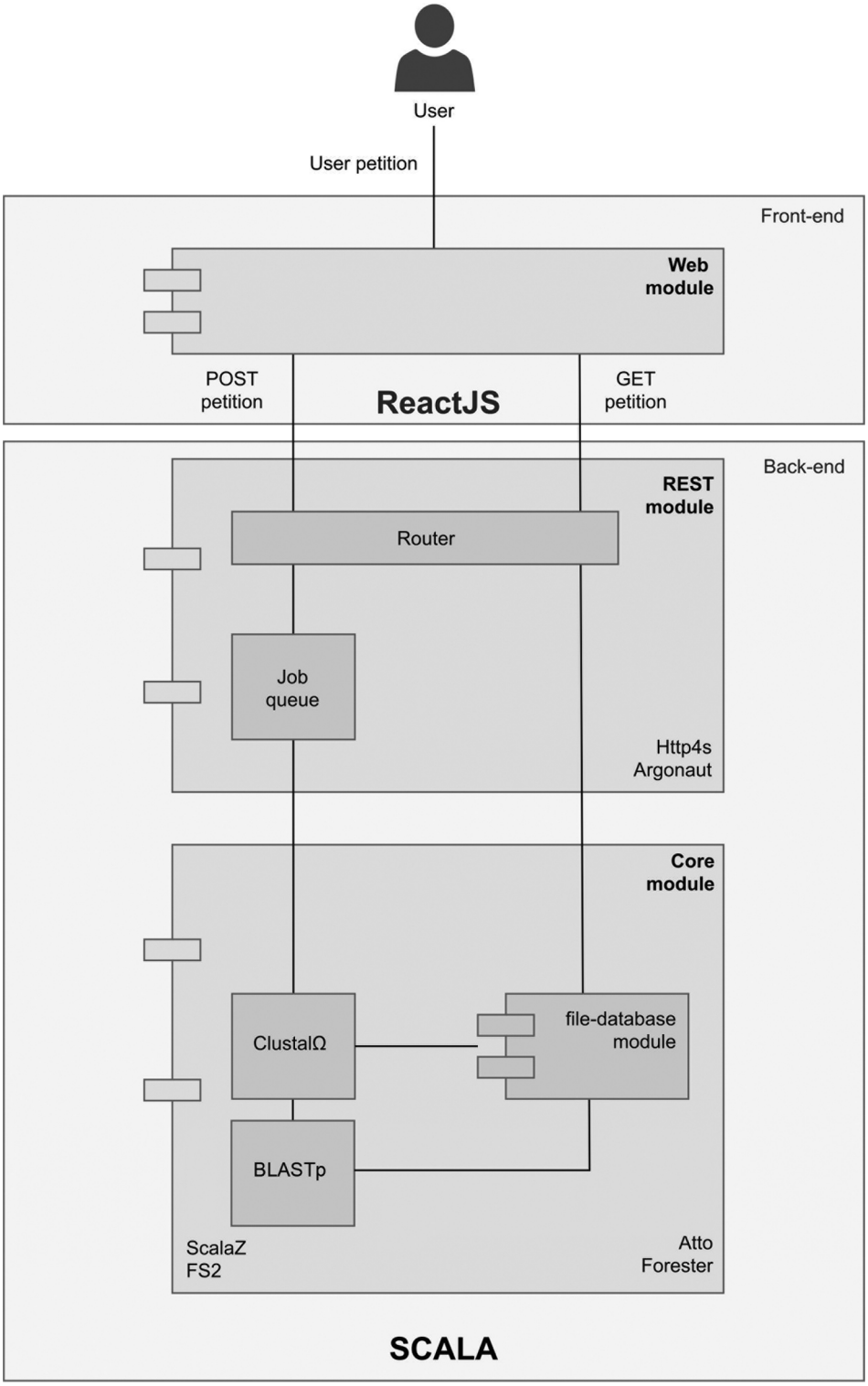


Fig. 2. The Scala-oriented modular architecture of Serpent. The core and REST modules form the back-end. The core module encompasses the main processes of analysis, while the REST module manages user requests by implementing a job queue for the incoming analysis requests. The Web module implements the user interface using the ReactJS library.

features, i.e. to be able to deal with new requirements and eventually, integrate other technologies. The modules of the current version of Serpent are as follows (Fig. 2): the core module is responsible for the execution of the different stages of the analysis pipeline, including the sequence alignment; the HTTP server module provides RESTful access to the functionalities of the core; and, the web module enables the user interface. The core and REST modules form the back-end of Serpent while the HTTP server module implements the front-end. Moreover, within the core module there is a non-explicit submodule, i.e. the file-database module, which manages the various data files generated throughout the execution of the workflow as well as the final analysis reports.

The design and functionality of the modules and, most notably, third-party dependencies, are described in the next subsections.

2.2.1. Core module

As the name suggests, this is the main component of the workflow. It is programmed in Scala with the aid of the libraries Scalaz [19], FS2 (Functional Streams for Scala) [20] and Atto [21]. Scalaz facilitates a purely functional style of programming and provides all the necessary constructs in order to develop the program in a completely functional manner. FS2 adds support for streaming input/output capabilities with an emphasis in compositionality, expressiveness, resource-safety and speed. FS2 can be seen as a replacement of old iterate-style constructs. Finally, Atto is applied to construct the data parsers applied at the different stages of the analysis.

Regarding third-party software, the core of Serpent relies on Clustal, the well-known and widely used multiple sequence alignment tool. Serpent integrates the latest version of the Clustal family, i.e. ClustalΩ, which offers increased stability and scalability [22]. Since the size of the proteomes under analysis may be considerable, the memory consumption of this tool needs to be controlled. To avoid excessive memory consumption, Serpent splits the set of proteomes under analysis into smaller, more manageable subsets, which can be processed by ClustalΩ without causing any issues. When all subsets are processed, the enriched proteins are merged into a single set of results. Then, BLASTp is executed on the predicted proteins to generate a statistical estimation of the quality of the results. Following the implementation proposed by Carroll H. D et al [23], with the p-values retrieved from the BLASTp alignments (since e-value = p-value * m, being m the size of the reference dataset), Bonferroni [24], Holm [25], Hommel [26], Hochberg [27] and Benjamini and Hochberg [28] and Benjamini and Yekutieli [29] corrections are estimated using the podkat R package version 1.4.2 (https://www.rdocumentation.org/packages/podkat/versions/1.4.2). The first four corrections are designed to give strong control of the family-wise error rate (FWER) while the last two corrections describe the false discovery rate (FDR), i.e. the expected proportion of false discoveries amongst the predicted proteins.

The bioactivity prediction report encompasses a tabular text file that lists the alignment results ordered by enrichment probability and a phylogenetic tree representation. The reporting functionality of Serpent is supported by Forester, a collection of open source Java libraries specialised in phylogenetic and evolutionary biology research [30]. Moreover, the phylogenetic tree is generated in the commonly used Newick [31] and PhyloXML [32] data formats, which allows further analysis in phylogeny-specific tools.

Serpent also integrates a biomedical ontology annotator, which supports the manual annotation of the reference sets by the user. These user-defined ontology annotations are automatically included in the results set. The inclusion of these annotations is not mandatory and is obviously dependent on the information available for the reference proteins.

The file-database module is responsible for managing all the data files and reports generated throughout the analysis (Table 1). As the input and output of each process are plain text files, the files are organised into directories and identified by a Universally Unique Identifier, UUID.

Table 1
Data files and reports available for a complete analysis. If the analysis fails, only the first two files in the table are available.

Name	Description
reference.fasta	The reference dataset
comparing.fasta	The proteome analysed
alignment.fasta	The output file with the alignments produced by ClustalΩ
full_report.csv	The report with the predicted proteins
similar.newick	The phylogenetic tree in Newick format
similar.phylo.xml	The phylogenetic tree in PhyloXML format
similar.png	The phylogenetic tree in PNG format

2.2.2. REST module

The HTTP Server module can be run directly without the assistance of a user interface. The API communication is based on the interchange of JSON objects. In particular, the analysis queue manager deals with the requests of analysis received by the server. This submodule implements a persistent job queue with First In, First Out (FIFO) ordering, i.e. the requests of analysis are served in the same order as they are issued. If unexpectedly the system crashes, the queue is restored to its last state upon restart. An ongoing analysis cannot be restarted to the exact point of the crash, but it is restarted from the last completed stage.

The Serpent server is programed in Scala making use of Http4s [33] and Argonaut [34] libraries. Http4s is a minimal, idiomatic, functional and completely asynchronous interface for HTTP services. In turn, Argonaut supports the interchange of JSON objects in server communication. The public REST API of Serpent, which is described in Table 2, is available at http://sing-group.org/serpent/api/.

2.2.3. Web module

Serpent also has a Web front-end, which was developed with commonly available Web technologies, i.e. HTML, CSS and JavaScript. Specifically, this interface leverages some of the state-of-the-art web capabilities offered by HTML5 (http://www.w3.org/TR/html5/) and CSS3 technologies (http://www.css3.info/) to create a visually appealing interface while retaining high usability.

The Javascript library ReactJS [35] is applied to maintain a reactive data-flow style, i.e. to abstract the document model of the HTML code and to enable a component-based programming environment. Moreover, the amCharts library is used to generate charts [36] while the Foundation front-end framework [37] is used to create the different components of the Web interface. Overall, the Web presentation of a bioactivity prediction report entails the description of the identified proteins in tabular format, an interactive phylogenetic tree viewer developed with react-phylocanvas, and a heatmap chart [36].

2.3. Performance and false discovery rate

The performance of the proposed method was assessed using four well-established metrics, i.e. precision ($\frac{TP}{TP+FP}$), recall ($\frac{TP}{TP+FN}$), accuracy ($\frac{TP+TN}{TP+TN+FP+FN}$) and FDR ($\frac{FP}{TP+FP}$) [38,39]. Two test cases supported this analysis, addressing the effect of the size of the reference dataset and the similarity threshold on the quality of the results.

The first test case measured the performance of the method while incrementing the size of the reference dataset. Therefore, the different glycoside hydrolases families related to 8 *Bacteroides* strains were retrieved from the CAZy database [40] (Supplementary Material S1). Three reference datasets were selected randomly for this test, having 8, 15 and 30 glycoside hydrolases from different families, respectively. Considering the phylogenetic distance between strains, a similarity threshold of 40% was selected. A “true positive” result was defined as a predicted family that was related to the strain according to the manually curated records of CAZy. These performance results can be found in the Supplementary Material S1.

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jbi.2019.103121.

The second test case assessed the performance while decreasing the similarity threshold value. Here, the protein sequences of 49 strains of *Staphylococcus aureus* were retrieved from the NCBI public FTP site (ftp://ftp.ncbi.nlm.nih.gov/) [41]. Protein entries that have not yet been annotated (i.e. hypothetical proteins) were not considered for analysis. The resulting dataset was tested against two isoforms of the methicillin resistance protein MecA considering similarity threshold values of 60%, 50%, 40% and 30%. A “true positive” result was defined as a predicted protein that was annotated as MecA in the GenBank database [42]. These performance results can be found in the Supplementary Material S2.

Table 2
The public REST API methods of Serpent.

Method	URL	Purpose
GET	project/*{uuid}*	Returns the information of the project, i.e. the UUID of each analysed proteome as well as the names of the reference set and the file keeping the list of proteomes in JSON format
GET	analysis/*{uuid}*	Returns basic information about the analysis, i.e. the alignment threshold, the execution status and the annotations in JSON format
GET	analysis/*{uuid}*/*{file}*	Returns the selected data file. The available files for a completed analysis, i.e. the reference dataset, the proteome analysed, the output file from ClustalΩ alignment, the prediction report and the phylogenetic tree in Newick, PhyloXML and PNG format (as described in Table 1)
GET	analysis/queue	Returns the size of the analysis queue, i.e. how many jobs are in line to be served, in JSON format
GET	analysis/queue/*{uuid}*	Returns the position of the proteome in the analysis queue in JSON format

3. Results

It is well known that most of antibiotic-resistance and foodborne diseases, which represent a global health burden, can be identified by at least one microbial protein, such as shiga and cholera toxins or antibiotic-resistance proteins [16,43]. So, the method described in this work holds great potential to deliver interesting results in translational clinical or biotechnological studies.

This rationale is demonstrated through the discussion of several cases studies involving metaproteomes as well as proteomes. The ability to predict taxon-specific functionality in metaproteomes is demonstrated using a case study related to the presence of prophages from *Clostridium* sp., *Escherichia coli, Ruminococcus bromii, Lactobacillus reuteri* and *Prevotella ruminicola* (Supplementary Material S3) and a case study describing the presence of β-lactam and methicillin resistance protein MecA (Supplementary Material S4). These are applications of biotechnological (phage-therapy) and clinical (antibiotic-resistance) interest, respectively. Additionally, we present two proteome case studies relevant to human gut microbiota research: the screening of a tetracycline resistance protein Tet(W) and the screening of *Bacteroides fragilis* and *Lactobacillus acidophilus* glycoside hydrolase repertoire. All this information is publicly available at the Serpent web service and its performance parameters including execution times and proteome/metaproteome(s) have been collected in the Supplementary Material S5.

The performance of the proposed method, and more specifically the FDR, is illustrated with two additional test cases, i.e. the screening of the glycoside hydrolases families of 8 *Bacteroides* strains (Supplementary Material S1), and the detection of the methicillin resistance protein MecA in 49 *Staphylococcus aureus* strains (Supplementary Material S2).

In general, all these analyses consisted of the following steps: (i) introduction of the details of the project, including the name, the target and reference datasets, the similarity threshold in percentage and the reference annotations (Fig. 3A), (ii) processing all the individual proteomes of the project (Fig. 3B), and (iii) generation of the bioactivity prediction results (Fig. 3C). To make visual inspection easier, results are summarised in two heatmaps: a heatmap depicting similarity percentages and another one representing alignment matches. For each analysed proteome, it is possible to check the detailed status and explore/download the reports generated during the analysis. All data can be downloaded. Most notably, heatmaps can be downloaded both as image (JPEG, PNG, SVG and PDF format) and as a table (CSV, XLSX and JSON format). The reference dataset and the list of analysed proteomes are also available for download.

3.1. Case study I: tetracycline resistance protein Tet(W)

This case study looked into the abundance of the Tet(W) protein into two datasets of different size: the *Bacteroides* genus (8 proteomes) and the larger *Lactobacillaceae* family (60 proteomes). The reference dataset was a compilation of different Tet(W) protein sequences present in intestinal bacteria. In a first attempt to analyse the proteomes of the *Bacteroides* genus dataset, an alignment threshold of 50% was set, but the absence of results led us to empirically lower this value to 30%. As

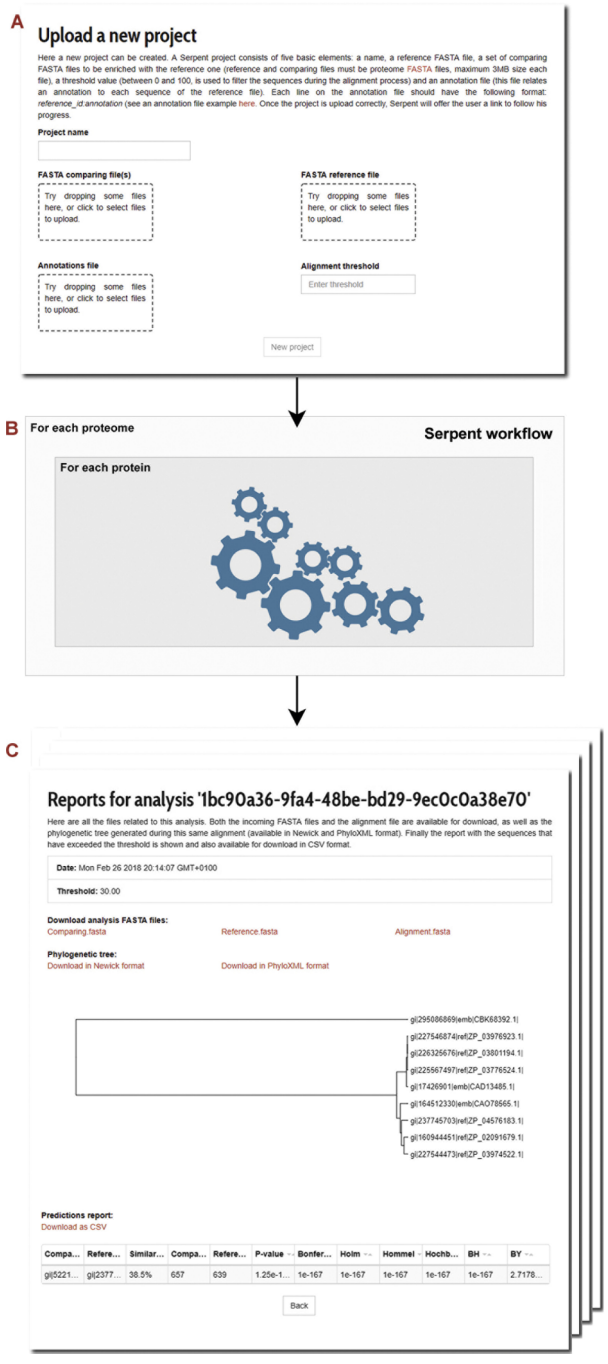
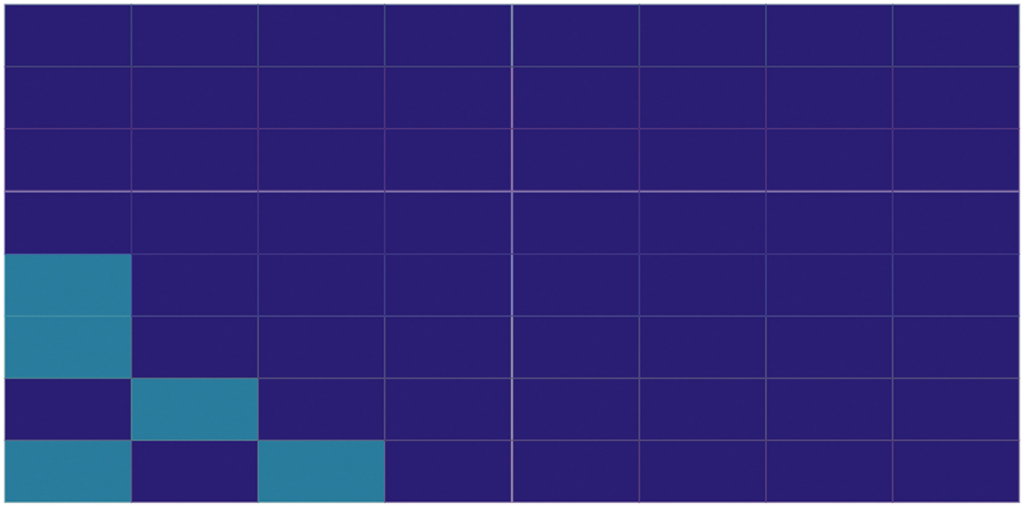


Fig. 3. Execution of a bioactivity potential prediction project in the Serpent web service. A new project consists in the introduction of the reference proteome, the proteomes to be analysed, the alignment threshold and the reference annotations (A). Proteomes are analysed individually (B). Phylogenetic trees and the enrichment report tables support expert analysis (C).

Similarity percentage heatmap: A



***Bacteroides fragilis* 638R:** B

Compa...	Refere...	Similar...	Compa...	Refere...	P-value	Bonfer...	Holm	Hommel	Hochb...	BH	BY
gi|5221...	gi|2377...	38.5%	657	639	1.25e-1...	1e-167	1e-167	1e-167	1e-167	1e-167	2.7178...

***Bacteroides fragilis* YCH46:** C

Compa...	Refere...	Similar...	Compa...	Refere...	P-value	Bonfer...	Holm	Hommel	Hochb...	BH	BY
gi|3011...	gi|2377...	30%	30	639	1	1	1	1	1	1	1

***Bacteroides salanitronis* DSM 18170:** D

Compa...	Refere...	Similar...	Compa...	Refere...	P-value	Bonfer...	Holm	Hommel	Hochb...	BH	BY
gi|3243...	gi|2377...	38.97%	641	639	5e-171	4e-170	4e-170	3.5e-170	4e-170	3e-170	8.1535...
gi|3243...	gi|2275...	30%	30	639	0.00975	0.078	0.078	0.04875	0.04875	0.0195	0.0529...

***Bacteroides xylanisolvens* XB1A:** E

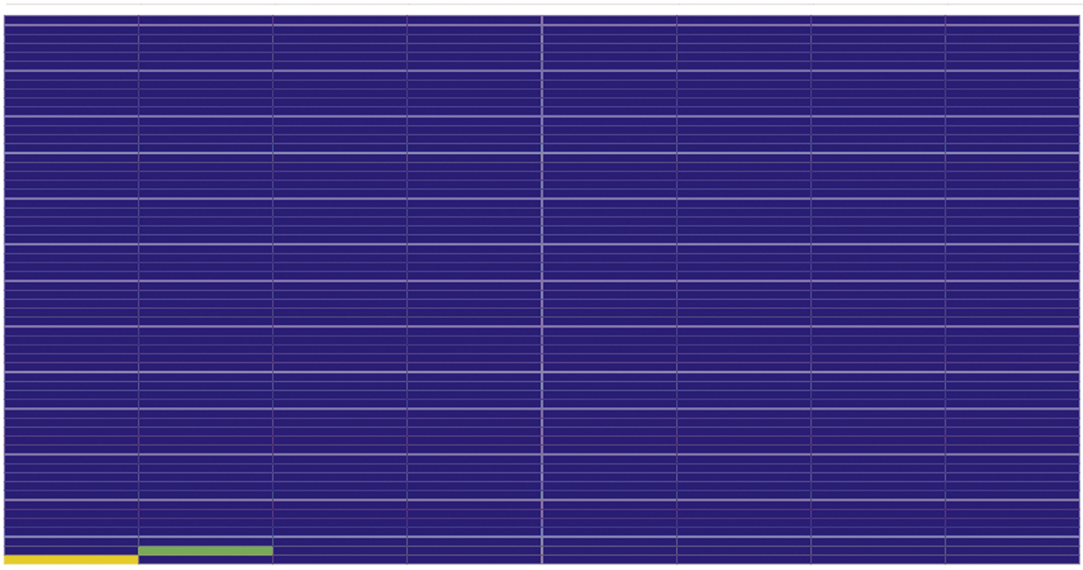
Compa...	Refere...	Similar...	Compa...	Refere...	P-value	Bonfer...	Holm	Hommel	Hochb...	BH	BY
gi|2950...	gi|2275...	38.5%	641	639	8.75e-1...	7e-168	7e-168	7e-168	7e-168	7e-168	1.9025...

Fig. 4. Prediction results for the Tetracycline resistance study of *Bacteroides* genus. Half of the strains showed Tet(W)-alike proteins. Similarity percentage values are coloured from blue (low) to yellow (high). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

illustrated in the Fig. 4, Serpent discovered Tet(W)-like proteins in four out of the eight strains of the *Bacteroides* dataset, i.e. *B. fragilis* 638R (Fig. 4B), *B. fragilis* YCH46 (Fig. 4C), *B. salanitronis* DSM 18170 (Fig. 4D) and *B. sylanisolvens* XB1A (Fig. 4E). However, looking at the statistical values of these predictions, two of these results, namely the proteins returned for *B. fragilis* 638R and *B. salanitronis* DSM 18170, were likely false positives due to their high adjusted p-values. This suspicion was later confirmed by domain experts. These false positives originated from the presence of poorly annotated proteins in the proteomes/metaproteomes of interest, including truncated proteins. This sort of situation is likely to occur when running Serpent with threshold values lower than 40–50%, as proteins with a low number of amino acids can be outputted. Therefore, a pre-processing filter was implemented for the deletion of such proteins from the input proteomes/metaproteomes. After discarding these predictions, Serpent predicted that three out of the eight strains (i.e. 42.86% of the strains) presented proteins similar to Tet(W), all with acceptable adjusted p-values (values very close to 0). Detailed information about these predictions is available in Supplementary Material S6 and the entire project is publicly available at http://sing-group.org/serpent/project/4cb5e624-1b21-11e8-9c54-0db66d62e496.

Regarding the *Lactobacillaceae* family, only two of the sixty strains showed Tet(W)-like proteins (Fig. 5), namely the strains *L. acidophilus* 30SC (Fig. 5B) and L. *casei* (Fig. 5C). Moreover, as in the case of *Bacteroides*, the results of *L. casei* contained non-curated, short protein sequences, resulting in false positives with p-values equal to 1. The obtained results showcase the lower abundance of Tet(W)-like proteins in this set of proteomes. Specifically, Tet(W)-like proteins were only discovered in one strain. Notably, the protein "gi|325334111|gb|ADZ08019.1|translation elongation factor G" from L. *acidophilus*

Similarity percentage heatmap: A



***Lactobacillus acidophilus* 30SC:** B

Compa...	Refere...	Similar...	Compa...	Refere...	P-value	Bonfer...	Holm	Hommel	Hochb...	BH	BY
gi|3253...	gi|1609...	99.22%	640	639	0	0	0	0	0	0	0

Lactobacillus casei: C

Compa...	Refere...	Similar...	Compa...	Refere...	P-value	Bonfer...	Holm	Hommel	Hochb...	BH	BY
gi|1907...	gi|2275...	100%	2	639	1	1	1	1	1	1	1
gi|1907...	gi|2275...	66.67%	3	639	1	1	1	1	1	1	1
gi|1907...	gi|2275...	60%	5	639	1	1	1	1	1	1	1
gi|1907...	gi|2275...	50%	6	639	1	1	1	1	1	1	1
gi|1907...	gi|2275...	50%	4	639	1	1	1	1	1	1	1
gi|1907...	gi|2275...	50%	2	639	1	1	1	1	1	1	1
gi|1907...	gi|2275...	50%	8	639	1	1	1	1	1	1	1
gi|1907...	gi|2275...	50%	8	639	1	1	1	1	1	1	1

Fig. 5. Prediction results for the Tetracycline resistance study of *Lactobacillae* family. Only *L. acidophilus* 30SC (B) and *L. casei* (C) showed Tet(W)-like proteins (A). Similarity percentage values are coloured from blue (low) to yellow (high). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

30SC shows a similarity of 99.22% with the *Faecalibacterium prausnitzii* Tet(W) protein (Fig. 5B).

Detailed information on these predictions is available in Supplementary Material S6 and the entire project is publicly available at http://sing-group.org/serpent/project/70fc0129-1c68-11e8-9c54-0db66d62e496.

3.2. Case study II: glycoside hydrolases

In this case study, thirty of the glycoside hydrolases encoded by the genome of the strain *Bacteroides fragilis* YCH46 and the thirty four glycoside hydrolases encoded by the strain *Lactobacillus acidophilus* 30SC were used as reference datasets for the prediction of bioactivity in the two proteome datasets presented in the case study I. Regarding the *Bacteroides* genus, and with the exception of *Bacteroides* sp. CF50 strain, all of the other species had predicted glycoside hydrolase proteins (Fig. 6). As general trend, Serpent predicted less glycoside hydrolase proteins while moving phylogenetically away from the strain *B. fragilis* YCH46, the strain from which the reference dataset was selected. As expected, the strain *B. fragilis* 638R contained the proteome in which more glycoside hydrolase proteins were retrieved, and most similarity values were over 99%, with only two exceptions. The other four strains presented similar values, with glycoside hydrolase protein identification percentages ranging between 40% and 60%. The entire project is publicly available at http://sing-group.org/serpent/project/26920afc-1d7e-11e8-9c54-0db66d62e496.

Regarding the *Lactobacillaceae* family, and as it is illustrated in Fig. 7, Serpent found glycoside hydrolase sequences in all the strains. The strains of *L. acidophilus* as well as those of *L. amylovorus* achieved the highest prediction percentages, with 85.29% of the reference glycoside hydrolase proteins identified, almost all of them with over 80% similarity. *L. crispatus* also had a high prediction rate, with 75.53% of

Analysis 'bb38f654-c527-4cc1-a046-06c5c34080ae**'**
Status: finished
Comparing File: Bacteroides_thetaiotaomicron_VPI-5482_uid399.fasta

Analysis '7880bf6d-8af5-453b-bb2d-f70f8071b6e8**'**
Status: finished
Comparing File: Bacteroides_xylanisolvens_XB1A_uid39177.fasta

Analysis '83d1d0c3-3546-4eb2-ac37-6479117768a6**'**
Status: finished
Comparing File: Bacteroides_fragilis_638R_uid50405.fasta

Analysis '9b899aad-368f-45a7-85e9-93639d29fccc**'**
Status: finished
Comparing File: Bacteroides_fragilis_YCH46_uid13067.fasta

Analysis 'cd45d2ef-e761-4cb8-8371-accc97540f53**'**
Status: finished
Comparing File: Bacteroides_vulgatus_ATCC_8482_uid13378.fasta

Analysis 'eb176b85-5b03-414c-944d-b5912fade9b7**'**
Status: finished
Comparing File: Bacteroides_salanitronis_DSM_18170_uid40055.fasta

Analysis '99f47e9d-9607-4be7-89c4-b73fb74415a2**'**
Status: finished
Comparing File: Bacteroides_helcogenes_P_36_108_uid41913.fasta

Analysis '47f265c7-9837-4be1-adaa-6fda77f95dc5**'**
Status: failed
Comparing File: Bacteroides_CF50_uid218492.fasta

Similarity percentage heatmap:

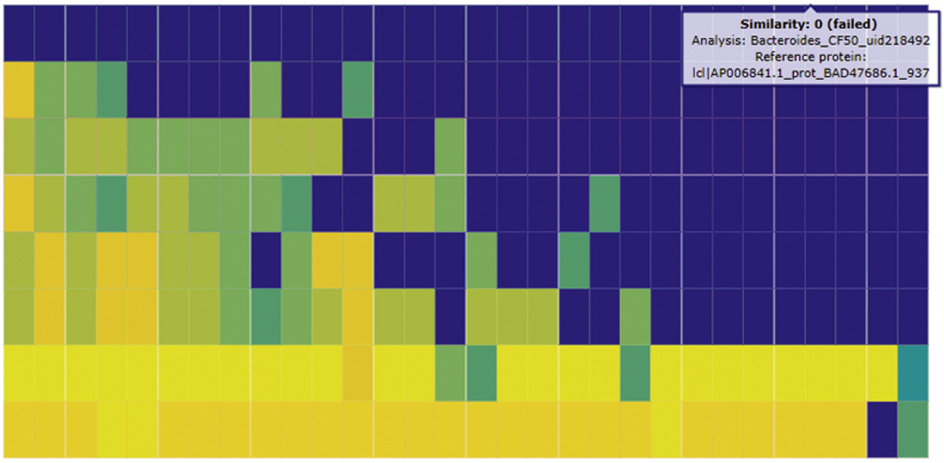


Fig. 6. Predictions obtained in the study of the glycoside hydrolase potential of *Bacteroides* genus. No glycoside hydrolase proteins were discovered in *Bacteroides* CF50. Similarity percentage values are coloured from blue (low) to yellow (high). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

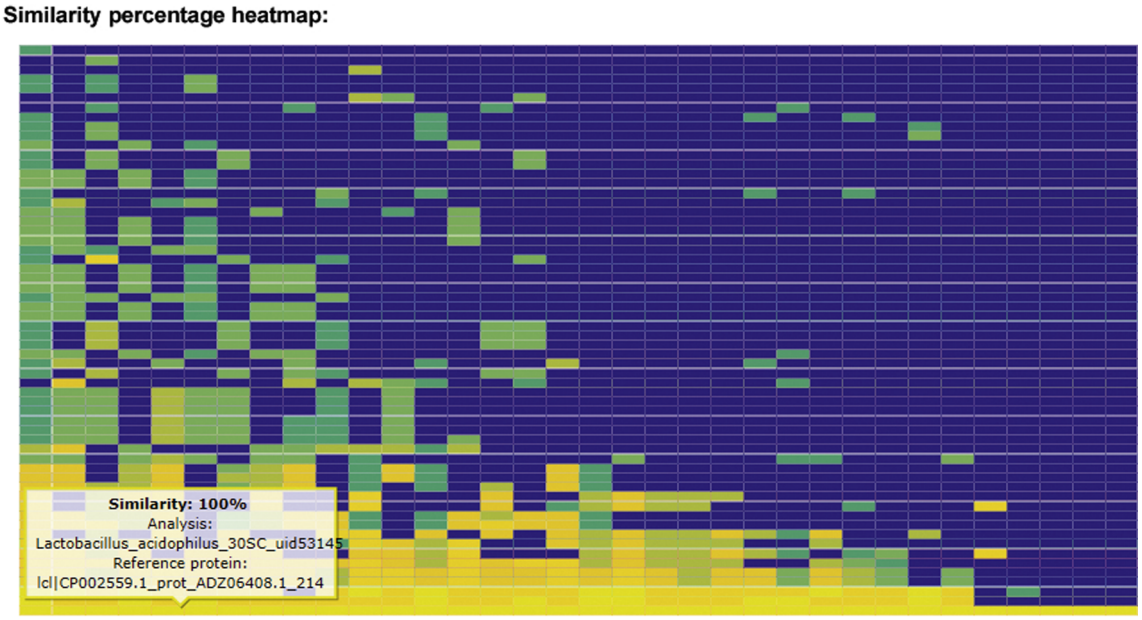


Fig. 7. Prediction results for the glycoside hydrolase study of *Lactobacillaceae* family. Serpent found glycoside hydrolase proteins in all the strains. Similarity percentage values are coloured from blue (low) to yellow (high). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the total number of glycoside hydrolases covered. As for the *Bacteroides* genus dataset, the other *Lactobacillaceae* strains presented fewer glycoside hydrolase predictions while moving taxonomically away from the *L. acidophilus* species, passing through *L. helveticus, L. jonshonni* and *L. kefiranofaciens*, until reaching *L. fermentum* and *L. pentosaceus* (which had only 2.94% of glycoside hydrolase proteins). The entire project is available at http://sing-group.org/serpent/project/11eb90bd-1d80-11e8-9c54-0db66d62e496.

4. Discussion

The method proposed here complements the classic functional annotation enrichment methods by enabling the discovery of functionally relevant proteins in proteomes or metaproteomes from sequence information and prior expert knowledge. Its goal is to allow the assignment of functions to hypothetical/putative/uncharacterised proteins whilst detecting specific bioactivities and recovering proteins from defined taxa. The practical use of such predictions may be to confirm a given assumption (e.g. check if some new isolate shares a characteristic bioactivity with the rest of the representatives from its family/species) or to investigate a novel hypothesis (e.g. to investigate the contribution of different genus/families and find out if they present a target bioactivity).

This analysis is scalable from a few up to hundreds of proteomes. In this sense, the new method allows the prediction of user-selected bioactivities in proteome and metaproteome datasets from different sources, e.g. strains, species or families, and the detection of precise protein sequences representing biomedical or biotechnologically relevant activities in hundreds of proteomes/metaproteomes at the time, facilitating the task to translational researchers.

In order to confirm the ability of Serpent for predicting bioactivity/functionality in proteomes, we have designed to two different experiments. A set of glycoside-hydrolases and tetracycline resistance proteins Tet(W) were investigated in two proteome collections of different size (8 vs 60 proteomes). Both cases were chosen given their direct relationship with the study of the human intestinal microbiota functionality and their timely and broad interest to the biomedical community [44]. In addition, these two collections of proteomes were chosen in an attempt to represent two different prediction patterns: in the case of glycoside hydrolases it was expected to find the bioactivity in most samples as the reference set is obtained from one of the genus representatives. On the contrary, the functionality represented by the Tet (W) protein was expected to have a minor presence as the set of reference includes sequences obtained from other microbial taxa.

For the first case study, eight TetW proteins were selected from the strains of representative Gram positive bacteria of the gut microbiota, namely *Bifidobacterium longum, Faecalibacterium prausnitzii, Lactobacillus reuteri, Coprococcus comes, Subdoligranulum variabile, Oxalobacter formigenes* and *Clostridium difficile* [45]. These eight proteins were used as reference set for the analysis of a dataset containing eight *Bacteroides* sp. proteomes and another dataset containing sixty *Lactobacillaceae* proteomes. *Bacteroides* genus includes species that are mainly commensal gut microbiota as well as opportunistic pathogens that, when clinically isolated, often show resistance to tetracycline [46]. In turn, the members of the *Lactobacillaceae* family are Gram positive, strict or aerotolerant anaerobic and fermentative rod-shaped bacteria that are commonly present among the microbiota of both human gastrointestinal and genitourinary tract, but also in fermented foods or on plant surfaces [47]. Some species of the genus *Lactobacillus* are also marketed as probiotics given their beneficial effects over human health [48].

Of the 38 tetracycline resistance genes already described, the gene *tet*(*W*), which encodes a ribosomal protection protein avoiding binding of tetracycline to ribosome, is the most extended tetracycline resistance gene in the anaerobic gut and rumen bacteria [49]. Tet(W) protein was mostly absent from the two proteome datasets, with very low similarity values in the case of *Bacteroides* genus, and just two out of the sixty proteomes of the *Lactobacillaeae* species containing Tet(W) homologues. Although it is well known that intestinal microbiota of humans and animals may act as a reservoir of antibiotic resistance genes that could ultimately be transferred to pathogens, Serpent analysis show that this is not a general trend in the two case studies presented [50]. This is very relevant from a biomedical point of view as transfer of antibiotic resistance genes is observed between bacterial species in the gastrointestinal tract of mammals, and for instance the Tet(W) protein of *Faecalibacterium prausnitzii* Tet(W) showed a high similarity value compared to the “translation elongation factor G” from *L. acidophilus* 30SC (99.22%) [50].

The low similarity of the Tet(W) proteins in our *Bacteroides* genus dataset contrasts, *a priori*, with previous experimental data showing that over 50% of the *Bacteroides fragilis* isolates are resistant to tetracyclines, while resistance in non-Bacteroides and other Gram-negative genera are more variable [51]. This can be explained by the fact that only tet(M), (Q) and (X) genes have been found in *Bacteroides* genus [52] whereas tet(W) gene is commonly found in other human and animal intestinal bacteria [53]. In this case, Serpent was again useful as after a BLASTp check, the predicted Tet(W)-like proteins were shown to be Tet(Q) proteins, another similar tetracycline resistance protein found in *Bacteroides* genus [52]. Noteworthy, low threshold values may return proteins with weak similarity, but this too can be a useful result, if the intention of the user is to explore sets of related proteins. Moreover, two of the Tet(Q) discovered proteins were annotated as "small GTP-binding protein". This shows the potential of Serpent to handle proteomes/metaproteomes with poor or even no annotations to predict their potential bioactivity, as it is possible to retrieve hypothetical or bad annotated proteins that have certain homology to the set of reference. This is a clear complement to classic experiments using functional enrichment algorithms such as SEA, GSEA or MEA.

Regarding the *Lactobacillaceae* family, the low number of sequences found in the dataset is in agreement with previous studies reporting the presence of tet(w) genes in few strains of *L. crispatus, L. johnsonii, L. paracasei* and *L. reuteri* [54–56]. Care should be taken in the sense that, in certain farming environments where antibiotic usage is frequent, the tet(W) gene can be present in more than 30% of the *Lactobacillus* isolates, mostly from the *L. ruminis*, *L. fermentum*, *L. reuteri* and *L. amylovorus* species [57].

Regarding the other case study, two sets of glycoside hydrolases from representatives of the *Bacteroides* genus and the *Lactobacillaceae* family were selected. The two datasets of reference were retrieved from the database of Carbohydrate-Active enZYmes (CAZy), which contains information about enzymes involved in the synthesis, metabolism, and transport of carbohydrates [58]. Specifically, a subset of thirty glycoside hydrolases of the *Bacteroides fragilis* YCH46 strain was used as reference dataset for the *Bacteroides* genus, and a dataset containing the thirty-four *Lactobacillus acidophilus* 30SC glycoside hydrolases served as reference dataset for the *Lactobacillaceae* family. Microbes inhabiting digestive tracts are exposed to diverse and abundant substrates, notably those derived from plant polysaccharides found in food or animal glycosaminoglycans produced by the host (e.g. mucins) [59–61]. The complete enzymatic deconstruction of polysaccharides involves not only glycoside hydrolases, but other CAZy enzymes such as polysaccharide lyases or carbohydrate esterases [62–65]. However, glycoside hydrolases are the most abundant enzymes involved in the breakdown of polysaccharides into simpler carbohydrates and they will therefore determine the ability to hydrolyse complex sugars [66].

Bacteroides genus are known to prosper in environments enriched in oligo and polysaccharides derived from plants, mainly due to the vast array of glycoside hydrolases encoded in their genomes [67,68]. On the contrary, members of the *Lactobacillae* family are mainly related to the hetero- or homo-fermentative conversion of simpler sugars into organic acids [69,70]. It is noteworthy that, in both cases, the glycolytic capacity is species-dependent as these enzymes reflect the adaptation to different animal guts (with different diets) or even different environments (dairy vs fermented meat, or gut vs vaginal mucosa) in the case of *Lactobacillus* genus. This is well illustrated by our analysis, where the number of glycoside hydrolase proteins is lower as the species is more taxonomically distant from the species included in the reference dataset (i.e. *B. fragilis* and *L. acidophilus*).

It is also noteworthy the added value of applying Serpent in translational studies. Many of the global health burdens correspond to microbial diseases in which the aetiological agent can be identified by at least one specific protein. In this regard, reference sets including a curated list of those protein biomarkers can be helpful for translational purposes, namely for studying antibiotic-resistance and foodborne diseases. Biomarkers for diarrheal diseases produced by cholera or Shiga-toxin producing *E. coli*, invasive enteric diseases produced by *Brucella* or *Listeria* spp. or even intoxications produced by *Clostridium botulinum* or *Bacillus cereus* could be detected using the proposed method. To demonstrate such translational application, two cases of studies supporting the screening of human faecal metaproteomes of clinical interest were designed: a case study related to the presence of prophages from specific taxa and a case study describing the presence of the antibiotic resistance protein MecA. In the case of the prophages study, given that phage therapy is currently becoming an alternative for the use of antibiotics, the massive identification of lysogenic phages within the intestinal microbiome has clinical relevance for the precise modification of microbial populations within the human gut microbiota. In the other case study, the MecA protein, usually present in *Staphylococcus aureus* genomes as well as in other bacteria, represents an important threat in clinical environments as it confers resistance to the semisynthetic penicillin methicillin.

On another level, it is important to stress that while the main aim of Serpent is to find bioactivities broadly, the method entails statistical mechanisms to help users to interpret and explore the results. Granted, any prediction tool has intrinsic the probability of making mistakes, that is, of producing false positives, and mismanagement of these false positives can lead to incorrect interpretations of the results. To demonstrate the ability of Serpent to manage false positives, two additional test cases were designed to evaluate the impact of the size of the reference dataset and the similarity threshold in the predictions. Most notably, the screening of the glycoside hydrolases families of 8 *Bacteroides* strains, and the detection of the methicillin resistance protein MecA in 49 *Staphylococcus aureus* strains.

In the first case, the larger the size of the reference dataset, i.e. the number of families of glycoside hydrolases, the greater the probability of false positives, i.e. predictions of such families (Supplementary Material S1). This is mostly due to the similarity between the reference sequences. In the case of the strains enriched with 30 families of glycoside hydrolases, the protein of the GH133 family presents a high false positive rate (in 62.5% of the strains). Such result is justified by the great similarity that this protein has with the protein of the GH5 family, which is present in all the strains in which G133 is falsely predicted. However, even with these false positives, the FDR is below 0.05 in the three executions (FDR of 0 with 8 families, 0.023 with 15 families and 0.046 with 30 families), which shows that, for the most part, Serpent is able to take into account these cases in sequence similarity analysis.

The second test case demonstrates the increase of the FDR with the decrease of the similarity threshold value. As detailed in the Supplementary Material S2, with intermediate threshold values, except for one protein in the *Staphylococcus aureus* strain LGA251, the probability of false positives was very low (FDR 0.026). Further examination of this false positive shows that the predicted protein was Pbp, a penicillin-binding protein with a 90% similarity with MecA (https://www.uniprot.org/uniprot/P07944). However, when the threshold was set under 40%, the rate of false positives increased greatly (for a similarity threshold of 30%, the FDR was 0.2). These false positives were proteins with a low number of aminoacids, predicted with high p-values (close to 1). This situation is similar to that previously observed in the analysis of the Tet(W) protein in *Bacteroides* and, as explained earlier, Serpent minimises its occurrence by enabling the filtering of such short proteins prior to the analysis.

5. Conclusions

This paper presents a new method for the prediction of the bioactivity potential of proteomes/metaproteomes based on prior knowledge. This methodology complements conventional functional annotation enrichment methods, allowing the assignment of functions to hypothetical/putative/uncharacterised proteins whilst detecting specific bioactivities or even recovering proteins from defined taxa.

Acknowledging that all tools of this nature have some embedded likelihood of making a mistake, the user is advised to take into consideration the FWER and FDR values calculated for the predictions as means to ensure a correct interpretation of the results.

The translational and practical use of the proposed method was demonstrated by two metaproteome examples of clinical interest and two proteome case studies relevant to the human gut microbiota research. Results show that the large information files obtained after DNA sequencing or protein identification experiments can be associated for translational analysis purposes that, in cases such as antibiotic-resistance pathogens or foodborne diseases, may introduce changes in how these important and global health burdens are approached in the clinical practice.

Conflicts of interest

Borja Sánchez is on the scientific board and is co-founder of Microviable Therapeutics SL. The other authors do not have competing interests.

Acknowledgment

This work was supported by the Spanish "Programa Estatal de Investigación, Desarrollo e Inovación Orientada a los Retos de la Sociedad" (grant AGL2013-44039R); the Asociación Española Contra el Cancer ("Obtención de péptidos bioactivos contra el Cáncer Colo-Rectal a partir de secuencias genéticas de microbiomas intestinales", grant PS-2016). This study was also supported by the Portuguese Foundation for Science and Technology (FCT) under the scope of the strategic funding of UID/BIO/04469/2013 unit and COMPETE 2020 (POCI-01-0145-FEDER006684). SING group thanks CITI (Centro de Investigación, Transferencia e Innovación) from University of Vigo for hosting its IT infrastructure.

References

[1] D.W. Huang, B.T. Sherman, R.A. Lempicki, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists, Nucleic Acids Res. 37 (2009) 1–13, https://doi.org/10.1093/nar/gkn923.

[2] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Gene Ontology: tool for the unification of biology, Nat. Genet. 25 (2000) 25–29, https://doi.org/10.1038/75556.

[3] U. Consortium, UniProt: a hub for protein information, Nucleic Acids Res. 43 (2015) D204–D212, https://doi.org/10.1093/nar/gku989.

[4] A. Fabregat, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, R. Haw, B. Jassal, F. Korninger, B. May, M. Milacic, C.D. Roca, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, G. Viteri, J. Weiser, G. Wu, L. Stein, H. Hermjakob, P. D'Eustachio, The Reactome pathway knowledgebase, Nucleic Acids Res. 46 (2018) D649–D655, https://doi.org/10.1093/nar/gkx1132.

[5] H. Tipney, L. Hunter, An introduction to effective use of enrichment analysis software, Hum. Genomics 4 (2010) 202, https://doi.org/10.1186/1479-7364-4-3-202.

[6] P. Khatri, S. Draghici, Ontological analysis of gene expression data: current tools, limitations, and open problems, Bioinformatics 21 (2005) 3587–3595, https://doi.org/10.1093/bioinformatics/bti565.

[7] R.K. Curtis, M. Orešič, A. Vidal-Puig, Pathways to the analysis of microarray data, Trends Biotechnol. 23 (2005) 429–435, https://doi.org/10.1016/j.tibtech.2005.05.011.

[8] P. Khatri, S. Draghici, G.C. Ostermeier, S.A. Krawetz, Profiling gene expression using onto-express, Genomics 79 (2002) 266–270, https://doi.org/10.1006/geno.2002.6698.

[9] S.W. Doniger, N. Salomonis, K.D. Dahlquist, K. Vranizan, S.C. Lawlor, B.R. Conklin, MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data, Genome Biol. 4 (2003) R7 <http://www.ncbi.nlm.nih.gov/pubmed/12540299>.

[10] B.R. Zeeberg, W. Feng, G. Wang, M.D. Wang, A.T. Fojo, M. Sunshine, S. Narasimhan, D.W. Kane, W.C. Reinhold, S. Lababidi, K.J. Bussey, J. Riss, J.C. Barrett, J.N. Weinstein, GoMiner: a resource for biological interpretation of genomic and proteomic data, Genome Biol. 4 (2003) R28 <http://www.ncbi.nlm.nih.gov/pubmed/12702209>.

[11] G. Dennis, B.T. Sherman, D.A. Hosack, J. Yang, W. Gao, H.C. Lane, R.A. Lempicki, DAVID: Database for Annotation, Visualization, and Integrated Discovery, Genome Biol. 4 (2003) P3 <http://www.ncbi.nlm.nih.gov/pubmed/12734009>.

[12] D.A. Hosack, G. Dennis, B.T. Sherman, H. Lane, R.A. Lempicki, Identifying biological themes within lists of genes with EASE, Genome Biol. 4 (2003) R70, https://doi.org/10.1186/gb-2003-4-10-r70.

[13] C.I. Castillo-Davis, D.L. Hartl, GeneMerge–post-genomic analysis, data mining, and hypothesis testing, Bioinformatics 19 (2003) 891–892 <http://www.ncbi.nlm.nih.gov/pubmed/12724301>.

[14] G.F. Berriz, O.D. King, B. Bryant, C. Sander, F.P. Roth, Characterizing gene sets with FuncAssociate, Bioinformatics 19 (2003) 2502–2504 <http://www.ncbi.nlm.nih.gov/pubmed/14668247>.

[15] A. Blanco-Míguez, A. Gutiérrez-Jácome, F. Fdez-Riverola, A. Lourenço, B. Sánchez, MAHMI database: a comprehensive MetaHitbased resource for the study of the mechanism of action of the human microbiota, Database. 2017 (2017), https://doi.org/10.1093/database/baw157.

[16] M.D. Kirk, S.M. Pires, R.E. Black, M. Caipo, J.A. Crump, B. Devleesschauwer, D. Döpfer, A. Fazil, C.L. Fischer-Walker, T. Hald, A.J. Hall, K.H. Keddy, R.J. Lake, C.F. Lanata, P.R. Torgerson, A.H. Havelaar, F.J. Angulo, World Health Organization estimates of the global and regional disease burden of 22 foodborne bacterial, protozoal, and viral diseases, 2010: a data synthesis, PLoS Med. 12 (2015) e1001921, https://doi.org/10.1371/journal.pmed.1001921.

[17] B.P. Blackburne, S. Whelan, Measuring the distance between multiple sequence alignments, Bioinformatics 28 (2012) 495–502, https://doi.org/10.1093/bioinformatics/btr701.

[18] M. Odersky, S. Micheloud, N. Mihaylov, M. Schinz, E. Stenman, M. Zenger, et al., An overview of the Scala programming language, 2004.

[19] Jetbrains, Scalaz, a Scala library for functional programming, n.d. < https://github.com/scalaz/scalaz > (accessed April 2, 2018).

[20] G. Coady, F.S. Thomas, M. Pilquist, FS2: Functional Streams for Scala, n.d. < https://github.com/functional-streams-for-scala/fs2 > (accessed April 2, 2018).

[21] E. Kmett, Atto, a compact, pure-functional, incremental text parsing library for Scala, n.d. < http://tpolecat.github.io/atto/ > (accessed April 2, 2018).

[22] F. Sievers, A. Wilm, D. Dineen, T.J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Soding, J.D. Thompson, D.G. Higgins, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, Mol. Syst. Biol. 7 (2014), https://doi.org/10.1038/msb.2011.75 539-539.

[23] H.D. Carroll, A.C. Williams, A.G. Davis, J.L. Spouge, Improving retrieval efficacy of homology searches using the false discovery rate, IEEE/ACM Trans. Comput. Biol. Bioinforma. 12 (n.d.) 531–7. http://doi.org/10.1109/TCBB.2014.2366112.

[24] C.E. Bonferroni, Teoria statistica delle classi e calcolo delle probabilità, Pubbl. Del R Ist. Super. Di Sci. Econ. e Commer. Di Firenze 8 (1936) 3–62.

[25] S. Holm, A simple sequentially rejective multiple test procedure, Scand. J. Stat. 6 (1979) 65–70 <https://www.jstor.org/stable/4615733>.

[26] G. Hommel, A stagewise rejective multiple test procedure based on a modified Bonferroni test, Biometrika 75 (1988) 383, https://doi.org/10.2307/2336190.

[27] Y. Hochberg, A sharper Bonferroni procedure for multiple tests of significance, Biometrika 75 (1988) 800, https://doi.org/10.2307/2336325.

[28] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. Ser. B 57 (1995) 289–300 <https://www.jstor.org/stable/2346101>.

[29] D. Yekutieli, Y. Benjamini, The control of the false discovery rate in multiple testing under dependency, Ann. Stat. 29 (2001) 1165–1188, https://doi.org/10.1214/aos/1013699998.

[30] C.M. Zmasek, Forester: software libraries for evolutionary biology and comparative genomics research, n.d. < https://sites.google.com/site/cmzmasek/home/software/forester > (accessed April 6, 2018).

[31] G. Olsen, The Newick tree format, n.d. < http://evolution.genetics.washington.edu/phylip/newicktree.html > (accessed February 14, 2018).

[32] M.V. Han, C.M. Zmasek, phyloXML: XML for evolutionary biology and comparative genomics, BMC Bioinf. 10 (2009) 356, https://doi.org/10.1186/1471-2105-10-356.

[33] Typelevel, Http4s, a minimal, idiomatic Scala interface for HTTP services, n.d. < https://http4s.org/ > (accessed April 2, 2018).

[34] T. Morris, S. Parsons, Argonaut, Purely Functional JSON in Scala, n.d. < http://argonaut.io/ > (accessed April 2, 2018).

[35] Facebook Inc., React - A JavaScript library for building user interfaces, 2018. < https://reactjs.org/ > (accessed February 14, 2018).

[36] AmCharts, AmCharts: JavaScript Charts & Maps, n.d. < https://www.amcharts.com/resources/ > (accessed April 6, 2018).

[37] ZURB Inc., Foundation: the most advanced responsive front-end framework in the world, n.d. < https://foundation.zurb.com/ > (accessed April 6, 2018).

[38] C. Yu, X. Li, H. Yang, Y. Li, W. Xue, Y. Chen, L. Tao, F. Zhu, Assessing the performances of protein function prediction algorithms from the perspectives of identification accuracy and false discovery rate, Int. J. Mol. Sci. 19 (2018) 183, https://doi.org/10.3390/ijms19010183.

[39] E. Ladoukakis, V. Pereira, E.G. Magny, A. Eyre-Walker, J. Couso, Hundreds of putatively functional small open reading frames in Drosophila, Genome Biol. 12 (2011) R118, https://doi.org/10.1186/gb-2011-12-11-r118.

[40] V. Lombard, H. Golaconda Ramulu, E. Drula, P.M. Coutinho, B. Henrissat, The carbohydrate-active enzymes database (CAZy) in 2013, Nucleic Acids Res. 42 (2014) D490-5, https://doi.org/10.1093/nar/gkt1178.

[41] NCBI Resource Coordinators, Database resources of the National Center for Biotechnology Information, Nucleic Acids Res. (2015), https://doi.org/10.1093/nar/gkv1290.

[42] D.A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, E.W. Sayers, GenBank, Nucleic Acids Res. 41 (2012) D36–D42, https://doi.org/10.1093/nar/gks1195.

[43] I. Vranakis, I. Goniotakis, A. Psaroulaki, V. Sandalakis, Y. Tselentis, K. Gevaert, G. Tsiotis, Proteome studies of bacterial antibiotic resistance mechanisms, J. Proteomics 97 (2014) 88–99, https://doi.org/10.1016/j.jprot.2013.10.027.

[44] J. Penders, E.E. Stobberingh, P.H.M. Savelkoul, P.F.G. Wolffs, The human microbiome as a reservoir of antimicrobial resistance, Front. Microbiol. 4 (2013), https://doi.org/10.3389/fmicb.2013.00087.

[45] J. Li, H. Jia, X. Cai, H. Zhong, Q. Feng, S. Sunagawa, M. Arumugam, J.R. Kultima, E. Prifti, T. Nielsen, A.S. Juncker, C. Manichanh, B. Chen, W. Zhang, F. Levenez, J. Wang, X. Xu, L. Xiao, S. Liang, D. Zhang, Z. Zhang, W. Chen, H. Zhao, J.Y. Al-Aama, S. Edris, H. Yang, J. Wang, T. Hansen, H.B. Nielsen, S. Brunak, K. Kristiansen, F. Guarner, O. Pedersen, J. Doré, S.D. Ehrlich, MetaHIT Consortium, P. Bork, J. Wang, MetaHIT Consortium, an integrated catalog of reference genes in the human gut microbiome, Nat. Biotechnol. 32 (2014) 834–841, https://doi.org/10.1038/nbt.2942.

[46] M.P. Nikolich, N.B. Shoemaker, A.A. Salyers, A Bacteroides tetracycline resistance gene represents a new class of ribosome protection tetracycline resistance, Antimicrob. Agents Chemother. 36 (1992) 1005–1012 1339256.

[47] P. Hütt, E. Lapp, J. Štšepetova, I. Smidt, H. Taelma, N. Borovkova, H. Oopkaup, A. Ahelik, T. Rööp, D. Hoidmets, K. Samuel, A. Salumets, R. Mändar, Characterisation of probiotic properties in human vaginal lactobacilli strains, Microb. Ecol. Heal. Dis. 27 (2016), https://doi.org/10.3402/mehd.v27.30484.

[48] C. Hill, F. Guarner, G. Reid, G.R. Gibson, D.J. Merenstein, B. Pot, L. Morelli, R.B. Canani, H.J. Flint, S. Salminen, P.C. Calder, M.E. Sanders, Expert consensus document: The International Scientific Association for Probiotics and Prebiotics consensus statement on the scope and appropriate use of the term probiotic, Nat. Rev. Gastroenterol. Hepatol. 11 (2014) 9, https://doi.org/10.1038/nrgastro.2014.66.

[49] M.S. Ammor, A.B. Florez, P. Alvarez-Martin, A. Margolles, B. Mayo, Analysis of tetracycline resistance tet(W) genes and their flanking sequences in intestinal Bifidobacterium species, J. Antimicrob. Chemother. 62 (2008) 688–693, https://doi.org/10.1093/jac/dkn280.

[50] A. Salyers, A. Gupta, Y. Wang, Human intestinal bacteria as reservoirs for antibiotic resistance genes, Trends Microbiol. 12 (2004) 412–416, https://doi.org/10.1016/j.tim.2004.07.004.

[51] H.M. Wexler, Anaerobic susceptibility testing: where are we and where do we go from here? Zentralbl. Bakteriol. 287 (1998) 1–5 <http://www.ncbi.nlm.nih.gov/pubmed/9532259>.

[52] M. Roberts, Acquired tetracycline and/or macrolide–lincosamides–streptogramin resistance in anaerobes, Anaerobe 9 (2003) 63–69, https://doi.org/10.1016/S1075-9964(03)00058-1.

[53] M. Egervärn, S. Roos, H. Lindmark, Identification and characterization of antibiotic resistance genes in *Lactobacillus reuteri* and *Lactobacillus plantarum*, J. Appl. Microbiol. 107 (2009) 1658–1668, https://doi.org/10.1111/j.1365-2672.2009.04352.x.

[54] S. Kastner, V. Perreten, H. Bleuler, G. Hugenschmidt, C. Lacroix, L. Meile, Antibiotic susceptibility patterns and resistance genes of starter cultures and probiotic bacteria used in food, Syst. Appl. Microbiol. 29 (2006) 145–155, https://doi.org/10.1016/j.syapm.2005.07.009.

[55] I. Klare, C. Konstabel, G. Werner, G. Huys, V. Vankerckhoven, G. Kahlmeter, B. Hildebrandt, S. Müller-Bertling, W. Witte, H. Goossens, Antimicrobial susceptibilities of Lactobacillus, Pediococcus and Lactococcus human isolates and cultures intended for probiotic or nutritional use, J. Antimicrob. Chemother. 59 (2007) 900–912, https://doi.org/10.1093/jac/dkm035.

[56] G. Huys, K. D'Haene, M. Danielsen, J. Mättö, M. Egervärn, P. Vandamme, Phenotypic and molecular assessment of antimicrobial resistance in lactobacillus paracasei strains of food origin, J. Food Prot. 71 (2008) 339–344, https://doi.org/10.4315/0362-028X-71.2.339.

[57] Y.-C. Chang, C.-Y. Tsai, C.-F. Lin, Y.-C. Wang, I.-K. Wang, T.-C. Chung, Characterization of tetracycline resistance lactobacilli isolated from swine intestines at western area of Taiwan, Anaerobe 17 (2011) 239–245, https://doi.org/10.1016/j.anaerobe.2011.08.001.

[58] B. Henrissat, G. Davies, Structural and sequence-based classification of glycoside hydrolases, Curr. Opin. Struct. Biol. 7 (1997) 637–644 <http://www.ncbi.nlm.nih.gov/pubmed/9345621>.

[59] B.D. Muegge, J. Kuczynski, D. Knights, J.C. Clemente, A. Gonzalez, L. Fontana, B. Henrissat, R. Knight, J.I. Gordon, Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans, Science (80-) 332 (2011) 970–974, https://doi.org/10.1126/science.1198719.

[60] L.A. David, C.F. Maurice, R.N. Carmody, D.B. Gootenberg, J.E. Button, B.E. Wolfe, A.V. Ling, A.S. Devlin, Y. Varma, M.A. Fischbach, S.B. Biddinger, R.J. Dutton, P.J. Turnbaugh, Diet rapidly and reproducibly alters the human gut microbiome, Nature 505 (2014) 559–563, https://doi.org/10.1038/nature12820.

[61] C.A. Lozupone, J.I. Stombaugh, J.I. Gordon, J.K. Jansson, R. Knight, Diversity, stability and resilience of the human gut microbiota, Nature 489 (2012) 220–230, https://doi.org/10.1038/nature11550.

[62] I. André, G. Potocki-Véronèse, S. Morel, P. Monsan, M. Remaud-Siméon, Sucrose-utilizing transglucosidases for biocatalysis, Top. Curr. Chem. 294 (2010) 25–48 <http://www.ncbi.nlm.nih.gov/pubmed/21626747>.

[63] G. Avigad, Enzymatic synthesis and characterization of a new trisaccharide, alpha-lactosyl-beta-fructofuranoside, J. Biol. Chem. 229 (1957) 121–129 <http://www.ncbi.nlm.nih.gov/pubmed/13491565>.

[64] J. Bircher, J. Müller, P. Guggenheim, U.P. Haemmerli, Treatment of chronic portal-systemic encephalopathy with lactulose, Lancet (London, England). 1 (1966) 890–892 <http://www.ncbi.nlm.nih.gov/pubmed/13491565>.

[65] N.F. Brás, P.A. Fernandes, M.J. Ramos, QM/MM studies on the β-galactosidase catalytic mechanism: hydrolysis and transglycosylation reactions, J. Chem. Theory Comput. 6 (2010) 421–433, https://doi.org/10.1021/ct900530f.

[66] R. Berlemont, A.C. Martiny, Glycoside hydrolases across environmental microbial communities, PLOS Comput. Biol. 12 (2016) e1005300, https://doi.org/10.1371/journal.pcbi.1005300.

[67] B.B. Matijašić, T. Obermajer, L. Lipoglavšek, I. Grabnar, G. Avguštin, I. Rogelj, Association of dietary type with fecal microbiota in vegetarians and omnivores in Slovenia, Eur. J. Nutr. 53 (2014) 1051–1064, https://doi.org/10.1007/s00394-013-0607-6.

[68] A. Benítez-Páez, E.M. Gómez del Pulgar, Y. Sanz, The glycolytic versatility of *Bacteroides uniformis* CECT 7771 and its genome response to oligo and polysaccharides, Front. Cell. Infect. Microbiol. 7 (2017), https://doi.org/10.3389/fcimb.2017.00383.

[69] O. Kandler, Carbohydrate metabolism in lactic acid bacteria, Antonie Van Leeuwenhoek 49 (1983) 209–224 6354079.

[70] J. Zheng, L. Ruan, M. Sun, M. Gänzle, A genomic view of Lactobacilli and Pediococci demonstrates that phylogeny matches ecology and physiology, Appl. Environ. Microbiol. 81 (2015) 7233–7243, https://doi.org/10.1128/AEM.02116-15.