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

Aitor Blanco-Míguez\textsuperscript{a,b,c}, Guillermo Blanco\textsuperscript{a,b,c}, Alberto Gutierrez-Jácome\textsuperscript{a}, Florentino Fdez-Riverola\textsuperscript{a,b,d}, Borja Sánchez\textsuperscript{c}, Anália Lourenço\textsuperscript{a,b,d,e,*}

\textsuperscript{a} ESEI: Escuela Superior de Ingeniería Informática, University of Vigo, Edificio Politécnico, Campus Universitario As Lagoas, s/n, 32004 Ourense, Spain
\textsuperscript{b} CINBIO - Centro de Investigaciones Biomédicas, University of Vigo, Campus Universitario Lagoas-Marcosende, 36310 Vigo, Spain
\textsuperscript{c} Department of Microbiology and Biochemistry of Dairy Products, Instituto de Productos Lácteos de Asturias (IPLA), Consejo Superior de Investigaciones Científicas (CSIC), Paseo Río Limarín, S/N, 33290 Villaviciosa, Asturias, Spain
\textsuperscript{d} SING Research Group, Galicia Sur Health Research Institute (IIS Galicia Sur), SERGAS-UVIGO, Hospital Álvaro Cunqueiro, 36312 Vigo, Spain
\textsuperscript{e} CEB - Centre of Biological Engineering, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal

\textbf{ARTICLE INFO}

\textbf{Keywords:}
Proteomes
Metaproteomes
Functionally relevant proteins
Bioactivity prediction
Translational application

\textbf{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 biological 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
1532-0464/ © 2019 Elsevier Inc. All rights reserved.
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 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
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. ClustalΩ2, 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 ClustalΩ2 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.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>reference.fasta</td>
<td>The reference dataset</td>
</tr>
<tr>
<td>comparing.fasta</td>
<td>The proteome analysed</td>
</tr>
<tr>
<td>alignment.fasta</td>
<td>The output file with the alignments produced by ClustalΩ2</td>
</tr>
<tr>
<td>full.report.csv</td>
<td>The report with the predicted proteins</td>
</tr>
<tr>
<td>similar.newick</td>
<td>The phylogenetic tree in Newick format</td>
</tr>
<tr>
<td>similar.phylo.xml</td>
<td>The phylogenetic tree in PhyloXML format</td>
</tr>
<tr>
<td>similar.png</td>
<td>The phylogenetic tree in PNG format</td>
</tr>
</tbody>
</table>

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 programmed 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, 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.
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 case 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%.

<table>
<thead>
<tr>
<th>Method</th>
<th>URL</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>GET project/uuid</td>
<td>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</td>
<td></td>
</tr>
<tr>
<td>GET analysis/uuid</td>
<td>Returns basic information about the analysis, i.e. the alignment threshold, the execution status and the annotations in JSON format</td>
<td></td>
</tr>
<tr>
<td>GET analysis/uuid/file</td>
<td>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)</td>
<td></td>
</tr>
<tr>
<td>GET analysis/queue</td>
<td>Returns the size of the analysis queue, i.e. how many jobs are in line to be served, in JSON format</td>
<td></td>
</tr>
<tr>
<td>GET analysis/queue/uuid</td>
<td>Returns the position of the proteome in the analysis queue in JSON format</td>
<td></td>
</tr>
</tbody>
</table>

Table 2
The public REST API methods of Serpent.
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. salantriosis* DSM 18170 (Fig. 4D) and *B. xylanisolvens* 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. salantriosis* 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](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*
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
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.)
the total number of glycoside hydrolases covered. As for the Bacteroides
genus dataset, the other Lactobacillaceae strains presented fewer gly-
coside hydrolase predictions while moving taxonomically away from
the L. acidophilus species, passing through L. helveticus, L. jonshornii 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 an-
notation enrichment methods by enabling the discovery of functionally
relevant proteins in proteomes or metaproteomes from sequence in-
formation and prior expert knowledge. Its goal is to allow the assign-
ment 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 bioac-
tivity).

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 rele-
vant 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 experi-
ments. 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 rela-
tionship with the study of the human intestinal microbiota function-
ality 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 re-
presentatives. On the contrary, the functionality represented by the Tet
(W) protein was expected to have a minor presence as the set of re-
ference 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 for-
migenes 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 commen-
sal 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 gastro-
intestinal 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 Lactobacillaceae 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 re-
sistance genes is observed between bacterial species in the gastro-
intestinal 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 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 glyco-side 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, glyco-side 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 Lactobacillaceae 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 amino acids, 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 Microvable Therapeutics SL. The other authors do not have competing interests.

Acknowledgment

This work was supported by the Spanish “Programa Estatal de Investigación, Desarrollo e Innovación Orientada a los Retos de la Sociedad” (grant AGL2013-44039R); the Asociación Española Contra el Cáncer (“Obtención de péptidos bioactivos contra el Cáncer Colo-Rectal”, grant PS-Cancer (“Obtención de péptidos bioactivos contra el Cáncer Colo-Rectal”); the Asociación Española Contrael Cáncer (“Obtención de péptidos bioactivos contra el Cáncer Colo-Rectal”, grant PS-Cancer (“Obtención de péptidos bioactivos contra el Cáncer Colo-Rectal”); the Asociación Española Contrael Cáncer (“Obtención de péptidos bioactivos contra el Cáncer Colo-Rectal”, grant PS-Cancer (“Obtención de péptidos bioactivos contra el Cáncer Colo-Rectal”). The translational and practical use of the proposed method was meansto ensure a correct interpretation of the results.

Liabilities

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.

References

I.Klare, C.Konstabel, G.Werner, G.Huys, V.Vankerckhoven, G.Kahlmeter, S.Kastner, V.Perreten, H.Bleuler, G.Hugenschmidt, C.Lacroix, L.Meile, A.A. Salyers, A.Gupta, Y.Wang, Human intestinal bacteria as reservoirs for antibiotic resistant genes that can be transferred to pathogens.

A.Salyers, A.Gupta, Y.Wang, Human intestinal bacteria as reservoirs for antibiotic resistant genes that can be transferred to pathogens.


P. Hütt, E. Lapp, J. Štšepetova, I. Smidt, H. Goossens, Antimicrobialsusceptibilitypatternsandresistancegenesofstarterculturesandprobioticbacteria are an area for further research.


P. Hütt, E. Lapp, J. Štšepetova, I. Smidt, H. Goossens, Antimicrobialsusceptibilitypatternsandresistancegenesofstarterculturesandprobioticbacteria are an area for further research.


P. Hütt, E. Lapp, J. Štšepetova, I. Smidt, H. Goossens, Antimicrobialsusceptibilitypatternsandresistancegenesofstarterculturesandprobioticbacteria are an area for further research.


P. Hütt, E. Lapp, J. Štšepetova, I. Smidt, H. Goossens, Antimicrobialsusceptibilitypatternsandresistancegenesofstarterculturesandprobioticbacteria are an area for further research.


P. Hütt, E. Lapp, J. Štšepetova, I. Smidt, H. Goossens, Antimicrobialsusceptibilitypatternsandresistancegenesofstarterculturesandprobioticbacteria are an area for further research.


P. Hütt, E. Lapp, J. Štšepetova, I. Smidt, H. Goossens, Antimicrobialsusceptibilitypatternsandresistancegenesofstarterculturesandprobioticbacteria are an area for further research.