PathogenFinder - Distinguishing Friend from Foe Using Bacterial Whole Genome Sequence Data.

Cosentino, Salvatore; Larsen, Mette Voldby; Aarestrup, Frank Møller; Lund, Ole

Published in: PloS one.

Link to article, DOI: 10.1371/journal.pone.0077302

Publication date: 2013

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
PathogenFinder - Distinguishing Friend from Foe Using Bacterial Whole Genome Sequence Data

Salvatore Cosentino1*, Mette Voldby Larsen1, Frank Møller Aarestrup2, Ole Lund1

1 Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kgs. Lyngby, Denmark, 2 National Food Institute, Technical University of Denmark, Kgs. Lyngby, Denmark

Abstract

Although the majority of bacteria are harmless or even beneficial to their host, others are highly virulent and can cause serious diseases, and even death. Due to the constantly decreasing cost of high-throughput sequencing there are now many completely sequenced genomes available from both human pathogenic and innocuous strains. The data can be used to identify gene families that correlate with pathogenicity and to develop tools to predict the pathogenicity of newly sequenced strains, investigations that previously were mainly done by means of more expensive and time consuming experimental approaches. We describe PathogenFinder (http://cge.cbs.dtu.dk/services/PathogenFinder/), a web-server for the prediction of bacterial pathogenicity by analysing the input proteome, genome, or raw reads provided by the user. The method relies on groups of proteins, created without regard to their annotated function or known involvement in pathogenicity. The method has been built to work with all taxonomic groups of bacteria and using the entire training-set, achieved an accuracy of 88.6% on an independent test-set, by correctly classifying 398 out of 449 completely sequenced bacteria. The approach here proposed is not biased on sets of genes known to be associated with pathogenicity, thus the approach could aid the discovery of novel pathogenicity factors. Furthermore the pathogenicity prediction web-server could be used to isolate the potential pathogenic features of both known and unknown strains.


Editor: Jason D. Barbour, University of Hawaii Manoa, United States of America

Received August 2, 2013; Accepted September 9, 2013; Published October 28, 2013

Copyright: © 2013 Cosentino et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by the Center for Genomic Epidemiology (http://www.genomicepidemiology.org) at the Technical University of Denmark and was funded by grant 09-067103/DSF from the Danish Council for Strategic Research. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: salvocos@cbs.dtu.dk

Introduction

Every year more than 15 millions deaths are the direct cause of infectious diseases, many of which are due to bacterial infections. Each year an estimated 1.3 million people die of tuberculosis and 0.2 millions of pertussis, while diarrhoea accounts for more than 2.5 millions deaths, and is one of the leading causes of death in worldwide [1]. But not all bacteria are dangerous and many of them are innocuous or even beneficial to human. The gut of a healthy adult human contains thousand of different microbial species, many of which are beneficial to their host, providing functions for nutrition, development, and regulating the immune response [2,3]. Nevertheless some bacterial species, like *Escherichia coli*, also include extremely deadly strains, causing for example diarrhoea, urinary tract infections, septicemia etc. Thus identifying pathogenic strains and understanding the biological mechanisms that cause them to become pathogenic is highly important to perform timely interventions and design control strategies, including interventions such as restrictions on contaminated food products, isolation of patients, correct treatment and development of targeted vaccines.

Ever since the 1880s the pathogenicity of bacteria have been assessed using Koch’s postulates, for human pathogens using animal models. During the last 2 decades many discoveries have shown that Koch’s postulates are not enough to decide if a given bacteria is pathogenic or not. The existence of diseases caused by bacteria that cannot grow in pure culture medium [4,5], the discovery of polymicrobial diseases [6], the role of metagenomic microbiota in chronic diseases [7], and last but not least, the discovery of Horizontal Gene Transfer (HGT) responsible for the swapping of genetic material between bacteria [8] (regardless the pathogenicity), are all cases in which the postulates have shortcomings. Already during his work with *Vibrio cholerae* Robert Koch himself discovered the shortcomings of animal models for correctly identifying human-specific pathogens. Thus, the use of animal models is not always reliable in defining if a given bacteria is human pathogenic. Moreover, assessing the pathogenicity by means of animal models or epidemiological studies is both time-consuming and expensive.

Among the molecular features that a bacterium needs to infect and survive inside its host [9] are exotoxins, endotoxins, two components systems [10], adherence factors, secretion systems (I to IV type) [11], through which bacteria can inject their toxins into its hosts cells [12]. Plasmids, secretion systems, and antibiotic resistance genes are commonly present in both commensal and pathogenic strains, while toxins are usually only present in pathogenic strains. There are many databases available containing genes encoding toxins and virulence factors along with other genes traditionally associated with pathogenicity [13,14].

One of the ways to classify a bacterium as human pathogenic using bioinformatics was (and still sometimes is) to look for some of these features in the genome of the isolate under investigation.

Among the molecular features that a bacterium needs to infect and survive inside its host [9] are exotoxins, endotoxins, two components systems [10], adherence factors, secretion systems (I to IV type) [11], through which bacteria can inject their toxins into its hosts cells [12]. Plasmids, secretion systems, and antibiotic resistance genes are commonly present in both commensal and pathogenic strains, while toxins are usually only present in pathogenic strains. There are many databases available containing genes encoding toxins and virulence factors along with other genes traditionally associated with pathogenicity [13,14].

One of the ways to classify a bacterium as human pathogenic using bioinformatics was (and still sometimes is) to look for some of these features in the genome of the isolate under investigation.
Unluckily this approach is not always reliable, partly because of HGT, which causes these features to be exchanged among pathogenic and innocuous strains of the same [15][16] or different species, an exchange which has been proved by the high amount of these features found in genomic islands [17]. Aside from the features directly associated to pathogenicity, there are also virulence "lifestyle" genes, important for the bacteria to survive inside the host and evade its immune system response [18][19], and genes that are, for example, needed to activate other genes, which are important in the processes of pathogenesis, even though they do not directly determine virulence. All the issues related to the prediction of bacterial pathogenicity based on phylogeny has caused researchers to look for different solutions.

Table 1. Training, test data and model parameters.

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Training Set</th>
<th>Test Set</th>
<th>Model Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pathogenic</td>
<td>Non-pathogenic</td>
<td>Pathogenic</td>
</tr>
<tr>
<td>TM-Alphaproteobacteria</td>
<td>29</td>
<td>60</td>
<td>89</td>
</tr>
<tr>
<td>TM-Betaproteobacteria</td>
<td>26</td>
<td>26</td>
<td>52</td>
</tr>
<tr>
<td>TM-Epsilonproteobacteria</td>
<td>17</td>
<td>5</td>
<td>22</td>
</tr>
<tr>
<td>TM-Gammaproteobacteria</td>
<td>122</td>
<td>97</td>
<td>219</td>
</tr>
<tr>
<td>TM-Acinobacteria</td>
<td>27</td>
<td>44</td>
<td>71</td>
</tr>
<tr>
<td>TM-Bacteroidetes</td>
<td>7</td>
<td>12</td>
<td>19</td>
</tr>
<tr>
<td>TM-Firmicutes</td>
<td>98</td>
<td>87</td>
<td>185</td>
</tr>
<tr>
<td>TM-Tenericutes</td>
<td>6</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>COMPL</td>
<td>40</td>
<td>174</td>
<td>214</td>
</tr>
<tr>
<td>WDM</td>
<td>372</td>
<td>513</td>
<td>885</td>
</tr>
</tbody>
</table>

The development of whole genome sequencing may open novel ways of predicting pathogenicity in bacterial species. In 1995 the genomes of *Mycoplasma genitalium* and *Haemophilus influenzae* [20,21] were completely sequenced, and scientists started considering the possibility of studying the pathogenesis of bacteria based on their genome sequences [22]. This was the start of a revolution that has been continuing during the last decade with the advent of Second-Generation or Next-Generation Sequencing (NGS), leading to a continuous decrease in sequencing costs and a fast development of sequencing technologies. At present, many different high-throughput sequencing systems are available [23–25] and the number of completely sequenced bacteria amount to almost 2,400 including more than 1,800 that have been submitted to the International Nucleotide Sequence Database Collaboration (INSDC) (www.genomesonline.org, May 2013).

A few methods have been proposed which make use of Support Vector Machines (SVM), BLAST or other bioinformatics tools to search for pathogenic features [26,27] or predict bacterial pathogenicity [28] by searching in pre-computed databases of genes associated with pathogens. One shared aspect among these methods is the fact that they restrict their search to well known pathogenic features, missing out on the information that may be contained in the many genes with unknown function. Furthermore, the methods ignore genes that could be shared and specific among non-pathogenic organisms. When bacteria become pathogenic through HGT their lifestyle change and some of the genes may be inactivated or even lost to adapt to the new lifestyle [29,30]. These genes are still present in non-pathogenic bacteria and hence could be used, together with the genes associated to pathogenicity, to separate dangerous bacteria from harmless ones.

As an alternative to the above mentioned prediction methods, we here developed a novel approach, building on a previous study [31]. In this study we selected groups of genes which are frequently found either in human pathogenic bacteria or in the innocuous ones, and show that this is more effective than using global similarity. Since we did not make any pre-assumption on the genes contained in our training-sets, we are able to identify new proteins associated to pathogenicity and also features shared among non-pathogenic bacteria. Moreover, our hypothesis-free approach gave us the chance to build, together with a phylogenetic-independent model using all the organisms we have, more specific models.
grouping organisms at different taxonomic ranks to improve the predictions in species like *E. coli*, in which the high amount of shared genes among pathogenic and commensal strains makes it particularly difficult to predict. In this study the original approach [31] was, furthermore, extended from γ-proteobacteria to all species and extended to not only give a prediction, but also identify which genes predicted to be most significantly associated with (or important for) pathogenicity or non-pathogenicity. Thus, the method will not only provide a prediction of pathogenicity, but may also be useful for identifying novel putative pathogenicity genes, supporting further functional genomic studies.

The predictor has been implemented as a free to use web-service, called *PathogenFinder*, to which users can upload raw reads, obtained from different NGS sequencing platforms, as well as assembled genomes, and obtain a fast estimation of the pathogenic potential of the bacteria they are studying, as well as the identification of potentially pathogenic genes. *PathogenFinder* could be helpful in situations of possible bacterial outbreaks, in which a fast analysis of the unknown strain is important to save lives, and follows the direction modern clinical microbiology [32] and global epidemiology [33] are taken driven by the revolution brought by high throughput DNA sequencing technologies.  

**Results and Discussion**

**Overview on the Created Models**

In this work we developed a method for predicting the pathogenicity of novel bacteria. We did this by comparing the proteins of the strain under investigation to a protein family database (PFDB) composed of groups of proteins (protein families or PFs) that were either associated with pathogenic or non-pathogenic organisms. In the creation of the PFDB we used 885 complete bacterial genomes (Table S1), 372 of which were tagged as human pathogens and 513 as non-pathogens.

All the proteins encoded by the bacterial genomes were initially clustered, and significant clusters, in which the majority of the proteins originated from either pathogens or non-pathogens, were identified. The PFs were accordingly tagged as pathogenic or non-pathogenic and a weight (Z-score) was calculated for each of them (see Materials and Methods for further details). Eight models were built using bacteria belonging to the same phylum or class as the query organism and accordingly select the appropriate model to be performed. The predictor has been implemented as a free to use web-server called *PathogenFinder*, to which a user can upload either the raw reads or the complete or draft genome of the organism they want to assess the pathogenicity of. One of the 10 built models (Table S1).  

Given a query organism, based on the number and kind of PFs that the proteins of the query organism are similar to, a prediction on whether it is human pathogenic or non-pathogenic is performed. The predictor has been implemented as a free to use web-server called *PathogenFinder*, to which a user can upload either the raw reads or the complete or draft genome of the organism they want to assess the pathogenicity of. One of the 10 built models can be selected for the prediction, and if the user does not know which class or phylum the organism belongs to, the web-server will identify it automatically by predicting 16S genes, using *RNAmmer* [34], and accordingly select the appropriate model to be used for the prediction. Both the set of matches used for the prediction and the raw matches from *PathogenFinder* are downloadable. The latter is particularly useful, since it contains more information about pathogenicity than the standard server output.

---

**Table 3.** Top 10 ranking pathogenic protein families and annotated functions of their proteins for TM-Gammaproteobacteria model.

<table>
<thead>
<tr>
<th>RANK</th>
<th>Z-score</th>
<th>P</th>
<th>N</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9.134</td>
<td>77</td>
<td>8</td>
<td>N-acetylmannosamine kinase (TCS)</td>
</tr>
<tr>
<td>2</td>
<td>8.500</td>
<td>49</td>
<td>0</td>
<td>Fimbrial proteins</td>
</tr>
<tr>
<td>3</td>
<td>8.170</td>
<td>62</td>
<td>6</td>
<td>Sialic Acid Transporter</td>
</tr>
<tr>
<td>4</td>
<td>8.158</td>
<td>53</td>
<td>3</td>
<td>Transposition helper protein</td>
</tr>
<tr>
<td>5</td>
<td>8.023</td>
<td>62</td>
<td>7</td>
<td>Acetyltransferase, type III secretion proteins</td>
</tr>
<tr>
<td>6</td>
<td>8.023</td>
<td>62</td>
<td>7</td>
<td>Macrolide-specific efflux, membrane protein</td>
</tr>
<tr>
<td>7</td>
<td>8.023</td>
<td>62</td>
<td>7</td>
<td>Type II secretion proteins</td>
</tr>
<tr>
<td>8</td>
<td>7.922</td>
<td>69</td>
<td>10</td>
<td>Unknown function, possible membrane proteins</td>
</tr>
<tr>
<td>9</td>
<td>7.906</td>
<td>60</td>
<td>7</td>
<td>Unknown function</td>
</tr>
<tr>
<td>10</td>
<td>7.855</td>
<td>53</td>
<td>4</td>
<td>Cytochrome b$_{552}$</td>
</tr>
</tbody>
</table>

P and N columns contain the number of pathogenic and non-pathogenic organisms in the protein family respectively. doi:10.1371/journal.pone.0077302.t003

**Table 4.** Top 10 ranking non-pathogenic protein families and annotated functions of their proteins for TM-Gammaproteobacteria model.

<table>
<thead>
<tr>
<th>RANK</th>
<th>Z-score</th>
<th>P</th>
<th>N</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>−6.52</td>
<td>3</td>
<td>34</td>
<td>Protein-L-isoaspartate</td>
</tr>
<tr>
<td>2</td>
<td>−6.44</td>
<td>2</td>
<td>31</td>
<td>Thi/Pf1 domain protein</td>
</tr>
<tr>
<td>3</td>
<td>−6.43</td>
<td>6</td>
<td>40</td>
<td>Anthranilate synthase component I</td>
</tr>
<tr>
<td>4</td>
<td>−5.98</td>
<td>6</td>
<td>36</td>
<td>8-amino-7-oxononanoate synthase</td>
</tr>
<tr>
<td>5</td>
<td>−5.92</td>
<td>5</td>
<td>34</td>
<td>Unknown function, putative transcriptional regulator</td>
</tr>
<tr>
<td>6</td>
<td>−5.82</td>
<td>0</td>
<td>21</td>
<td>Adenosylmethionine decarboxylase</td>
</tr>
<tr>
<td>7</td>
<td>−5.81</td>
<td>8</td>
<td>39</td>
<td>Unknown function</td>
</tr>
<tr>
<td>8</td>
<td>−5.80</td>
<td>2</td>
<td>26</td>
<td>Unknown function, probable condensation protein</td>
</tr>
<tr>
<td>9</td>
<td>−5.68</td>
<td>0</td>
<td>20</td>
<td>Nitrite transporter</td>
</tr>
<tr>
<td>10</td>
<td>−5.62</td>
<td>1</td>
<td>22</td>
<td>Glucose-galactose transporter</td>
</tr>
</tbody>
</table>

P and N columns contain the number of pathogenic and non-pathogenic organisms in the protein family respectively. doi:10.1371/journal.pone.0077302.t004
and could hence be used for a more detailed analysis of the pathogenicity features of the organisms under investigation.

**Performance on Five-Fold Cross Validation and Independent Test Data**

The TM models were tested using only organisms belonging to the specific phylum/class, while in the case of the WDM model the whole independent data-set was used for the test.

Table 2 shows the performance of the ten models as obtained by 5-fold cross validation (CV) (column 2) and on independent test-sets of organisms from the same taxonomic group (column 3). As can be seen for the Tenericutes and Bacteroidetes phyla, the performances were very poor when compared to the MCC obtained in the CV tests. This is likely to be caused by the models being built using a small number of organisms (Table 1). Furthermore, it was tested on a set of organisms from species that were not present in the training-set.

To compare the performance of the WDM model to those of the TM and COMPL models, we examined the MCC of the WDM on the same test-sets used for the other models (column 4 in Table 2).

For example, to examine the performance of the WDM in predicting the pathogenicity of Firmicutes bacteria, we tested it with the same organisms used to assess the accuracy of the TM-Firmicutes model.

The MCC obtained by the WDM (0.758) on all bacteria was higher than the overall accuracy of all the TM models and COMPL model combined (0.736). Nonetheless, the TM models performed better for bacteroidetes, α, β, and γ-proteobacteria, even though for the latter the difference from the WDM was not significant. The remaining TM models and the COMPL model had lower MCC than the WDM for the same organisms.

**Performance on Draft Genomes and *Escherichia coli***

The models ability in predicting the pathogenicity of an isolate as based on a draft genome was tested using 259 sets of illumina raw reads from 6 different species. While in the case of *Campylobacter jejuni*, *Klebsiella pneumoniae* and *Staphylococcus aureus* (57 isolates in total) all the predictions were correct, the results were not satisfactory for Enterococci and *E. coli*. Of 50 Enterococcus

---

### Table 5. Top 10 ranking pathogenic protein families and annotated functions of their proteins for the WDM model.

<table>
<thead>
<tr>
<th>RANK</th>
<th>Z-score</th>
<th>P</th>
<th>N</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10.18</td>
<td>38</td>
<td>0</td>
<td>Borrelia Plasmid partition proteins</td>
</tr>
<tr>
<td>2</td>
<td>9.49</td>
<td>33</td>
<td>0</td>
<td>TCS associated genes, unknown functions</td>
</tr>
<tr>
<td>3</td>
<td>9.19</td>
<td>31</td>
<td>0</td>
<td>Lipoate-protein ligase, lipoate metabolism associated proteins</td>
</tr>
<tr>
<td>4</td>
<td>9.19</td>
<td>31</td>
<td>0</td>
<td>Unknown functions, flavin oxidoreductase</td>
</tr>
<tr>
<td>5</td>
<td>9.04</td>
<td>30</td>
<td>0</td>
<td>Exfoliative toxin A</td>
</tr>
<tr>
<td>6</td>
<td>8.89</td>
<td>29</td>
<td>0</td>
<td>Pili assembly proteins, Motility, Secretion Systems</td>
</tr>
<tr>
<td>7</td>
<td>8.89</td>
<td>30</td>
<td>0</td>
<td>Unknown function, shikimate kinase</td>
</tr>
<tr>
<td>8</td>
<td>8.89</td>
<td>29</td>
<td>0</td>
<td>Pili assembly proteins, Motility, Secretion Systems</td>
</tr>
<tr>
<td>9</td>
<td>8.74</td>
<td>28</td>
<td>0</td>
<td>Multiple antibiotic resistance (MarR) family proteins</td>
</tr>
<tr>
<td>10</td>
<td>8.74</td>
<td>28</td>
<td>0</td>
<td>Mutarotase Yjht (sialic acid mutarotation), unknown functions</td>
</tr>
</tbody>
</table>

P and N columns contain the number of pathogenic and non-pathogenic organisms in the protein family respectively.

doi:10.1371/journal.pone.0077302.t005

### Table 6. Top 10 ranking non-pathogenic protein families and annotated functions of their proteins for the WDM model.

<table>
<thead>
<tr>
<th>RANK</th>
<th>Z-score</th>
<th>P</th>
<th>N</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>−6.68</td>
<td>63</td>
<td>0</td>
<td>tRNA proteins</td>
</tr>
<tr>
<td>2</td>
<td>−6.62</td>
<td>62</td>
<td>0</td>
<td>ABC transporter related proteins (for δ and χ proteobacteria)</td>
</tr>
<tr>
<td>3</td>
<td>−6.18</td>
<td>54</td>
<td>0</td>
<td>Rubrerythrin</td>
</tr>
<tr>
<td>4</td>
<td>−6.07</td>
<td>52</td>
<td>0</td>
<td>Rubrerythrin</td>
</tr>
<tr>
<td>5</td>
<td>−6.01</td>
<td>51</td>
<td>0</td>
<td>Iron-sulfur binding domain proteins</td>
</tr>
<tr>
<td>6</td>
<td>−6.01</td>
<td>51</td>
<td>0</td>
<td>Hydroxymethylglutaryl-CoA synthase</td>
</tr>
<tr>
<td>7</td>
<td>−5.95</td>
<td>50</td>
<td>0</td>
<td>Unknown function</td>
</tr>
<tr>
<td>8</td>
<td>−5.89</td>
<td>49</td>
<td>0</td>
<td>Unknown function</td>
</tr>
<tr>
<td>9</td>
<td>−5.83</td>
<td>48</td>
<td>0</td>
<td>Unknown function</td>
</tr>
<tr>
<td>10</td>
<td>−5.70</td>
<td>46</td>
<td>0</td>
<td>Sulfite reductase subunit</td>
</tr>
</tbody>
</table>

P and N columns contain the number of pathogenic and non-pathogenic organisms in the protein family respectively.

doi:10.1371/journal.pone.0077302.t006
faecalis and 49 Enterococcus faecium from healthy Danish pigs, all isolates were predicted as pathogenic. Our training-set only contained a single pathogenic E. faecalis and no E. faecium, which may explain these results.

The WDM as well as the TM-Gammaproteobacteria models predicted the 10 E. coli strains in the test-set as pathogenic, although 4 strains were annotated as non-pathogenic. A similar situation was observed for the 103 E. coli draft genomes. Accordingly, we decided to create a model only for the Enterobacteriaceae family, using the organisms in our training-set. The resulting model correctly predicted 1 of 4 non-pathogenic E. coli achieving an MCC of 0.41, but all draft genomes were still predicted as pathogenic. The model also showed improvements in predicting other Enterobacteriaceae, with an MCC of 0.675, while WDM and TM-Gammaproteobacteria had an MCC of 0.519 and 0.617, respectively.

To improve the predictions for E. coli further, we decided to create 2 special models. These models were called ecoli_boost and enterobac_boost, and they were trained on a set that was enriched with 14 extra non-pathogenic E. coli strains downloaded from the National Center for Biotechnology Information (NCBI) (Table S2). These two models had a noticeable improvement on both CG test-sets and on the 103 assembled E. coli isolates, on which MCC was 0.346 (Acc = 67%) and 0.360 (Acc = 68%) for enterobac_boost and ecoli_boost, respectively. The lists of organisms used to train the enterobac_boost and ecoli_boost models, together with more details on the results on E. coli can be seen in Table S2.

Comparison to other Prediction Methods

Presently, the literature describes two main approaches for predicting the human pathogenicity of bacteria based on whole genome sequencing data: the first, proposed by Andreatta et al. [31], is able to predict the pathogenicity of x-proteobacteria, and it was from this study we borrowed the concept of PFs; the second method, developed by Iraola et al. [28], uses SVM [35], and can predict the pathogenicity of all types of bacteria. In this method the authors selected 120 genes associated to pathogenicity from 600 complete genomes using SVM, and built a prediction model based on the selected genes.
<table>
<thead>
<tr>
<th>Bacterial Group</th>
<th>Training Set</th>
<th>Test Set</th>
<th>Protein Families</th>
</tr>
</thead>
<tbody>
<tr>
<td>TM-Alphaproteobacteria</td>
<td>'&amp;&amp;$%&amp;</td>
<td>$&amp;&amp;$%&amp;</td>
<td>!&quot;#$%&amp;</td>
</tr>
<tr>
<td>TM-Betaproteobacteria</td>
<td>!#&amp;&amp;$&amp;</td>
<td>!#&amp;&amp;$&amp;</td>
<td>!#&amp;&amp;$&amp;</td>
</tr>
<tr>
<td>TM-Epsilonproteobacteria</td>
<td>$&amp;&amp;$%&amp;</td>
<td>$&amp;&amp;$%&amp;</td>
<td>$&amp;&amp;$%&amp;</td>
</tr>
<tr>
<td>TM-Gammaproteobacteria</td>
<td>$$&amp;&amp;$%&amp;</td>
<td>$$&amp;&amp;$%&amp;</td>
<td>$$&amp;&amp;$%&amp;</td>
</tr>
<tr>
<td>TM-Actinobacteria</td>
<td>)&amp;&amp;$%&amp;</td>
<td>)&amp;&amp;$%&amp;</td>
<td>)&amp;&amp;$%&amp;</td>
</tr>
<tr>
<td>TM-Bacteroidetes</td>
<td>)&amp;&amp;$%&amp;</td>
<td>)&amp;&amp;$%&amp;</td>
<td>)&amp;&amp;$%&amp;</td>
</tr>
<tr>
<td>TM-Firmicutes</td>
<td>)&amp;&amp;$%&amp;</td>
<td>)&amp;&amp;$%&amp;</td>
<td>)&amp;&amp;$%&amp;</td>
</tr>
<tr>
<td>TM-Tenericutes</td>
<td>)&amp;&amp;$%&amp;</td>
<td>)&amp;&amp;$%&amp;</td>
<td>)&amp;&amp;$%&amp;</td>
</tr>
</tbody>
</table>

**Tot. Protein Families**
- TM-Alphaproteobacteria: 32961
- TM-Betaproteobacteria: 17090
- TM-Epsilonproteobacteria: 1494
- TM-Gammaproteobacteria: 40537
- TM-Actinobacteria: 19018
- TM-Bacteroidetes: 9620
- TM-Firmicutes: 16392
- TM-Tenericutes: 1536
- COMPL: 67150
- WDM: 192411

**PFDB Size (No. Seq./Mbytes)**
- TM-Alphaproteobacteria: 179809/82,7
- TM-Betaproteobacteria: 76905/35,2
- TM-Epsilonproteobacteria: 13128/4,4
- TM-Gammaproteobacteria: 229030/100,6
- TM-Actinobacteria: 179809/29,6
- TM-Bacteroidetes: 33502/16,9
- TM-Firmicutes: 99754/41,7
- TM-Tenericutes: 3965/2,0
- COMPL: 311858/142,1
- WDM: 784316/346,5
To compare our method to the one proposed by Andreotta et al., we built a model using the same set of z-proteobacteria organisms (155) and the same parameters (MinOrg, HT, LT) used by Andreotta et al. The key differences between our method and the one by Andreotta et al. are: 1) we used CD-HIT instead of BLAST in both the protein clustering and prediction phases; 2) we used Equation 3 to filter the significant matches of the query sequences, while Andreotta et al. filtered based on a BLAST e-value threshold; 3) We compute the final predictions using the Z-scores, while Andreotta et al. counted the number of pathogenic and non-pathogenic families matched. The obtained model was tested on the same independent set used by Andreotta et al. This set included 24 organisms (14 pathogenic), and our model was able to correctly classify 23 organisms (95.8%). This is equivalent to an MCC of 0.92, while Andreotta’s MCC was 0.837. The one organism that our method was not able to correctly classify, is *Salmonella enterica* Serovar Gallinarum str. 287/91 [GenBank:30689], which is pathogenic for poultry, but not known to be for humans. The pathogenicity of this organism is restricted to chicken although it shares a high quantity of genomic features associated to pathogenicity with its human pathogenic ancestor *Salmonella Enteritidis* [36]. It is likely that these features mislead the prediction model, since also the method by Andreotta et al. wrongly classified this *S. enterica* strain.

To compare our method to the predictor proposed by Iraola et al., we used the test-set they used for their blind test evaluation. The test-set, originally composed of 233 organisms, contained 5 strains, which were excluded from the comparison, since they were also present in our training-set. Overall, for the comparison, we had a test-set composed of 228 organisms, 192 of which are tagged as human pathogens and the remaining 36 as non-pathogens. *PathogenFinder* achieved an overall MCC of 0.67 for the taxonomy models and 0.65 for the WDM model. Both results are higher than the MCC of 0.6 obtained by the method proposed by Iraola et al. Table S3 contains a detailed description of the comparison, including the organisms used and the corresponding predictions from both methods.

**PFDB Analysis and Biological Interpretation**

For each created model, an analysis of its PFDB was performed and its PFs ranked based on their Z-scores. The scores above 0 are associated with pathogenic PFs, while those below 0 are associated with non-pathogenic PFs. No protein function analysis was done prior to the models creation, making the approach unbiased on the genomic content of the organisms, regardless of their pathogenicity. In this paragraph we describe the analysis of the PFs of the TM-Gammaproteobacteria and WDM.

The analysis of the PFDB of TM-Gammaproteobacteria model showed that the high ranked pathogenic families (Table 3) contained proteins well known to be associated to pathogenicity. The family at rank 1 and 3 contained N-acetylglucosamine kinase, which is a key enzyme in sialic acid synthesis and sialic acid transport proteins. Sialic acid is important for virulence and is believed to help the microbes to disguise themselves as host cells in order to evade the host’s immune system response [37]. Fimbrial proteins (rank 2) are important for bacterial adherence [38]. At rank 10 we found cytochrome b$_{592}$ proteins that help bacteria to survive and grow in conditions of poor oxygen [39]. Other high-ranked families contained proteins associated with secretion systems (II and III) and antibiotic resistance.

An interesting finding, which was also found in [31], was the presence of families containing proteins with unknown functions associated with pathogenicity. This finding suggests that those proteins with unknown function might have important roles in the bacterial pathogenesis and could form the basis for further functional studies improving our understanding of bacterial pathogenicity. Proteins with unknown functions were also identified as associated with non-pathogenic PFs (Table 4).

The analysis of the PDBF of the WDM enabled us to see if proteins involved (or not involved) in pathogenesis belong to organisms of different taxonomy, and at the same time gave us an insight on how proteins are conserved along the different phyla. Again, we found that the top ranked families associated to pathogenicity (Table 5) contained also proteins with unknown function.

The highest ranked PF contained proteins encoded by plasmids from different pathogenic *Borelia* species (mainly *Burgdorferi*), which are involved in pathogenesis [40,41]. The family ranked 3rd contained proteins associated with lipoate metabolism. The acquisition and use of lipoate by pathogens affect their virulence and the pathogenesis of the diseases they cause [42]. Among the toxins found were: exoflagellin A (family-rank 5) in *Staphylococcus aureus* strains, causative of Staphylococcal scalded skin syndrome [43,44]; streptolysin O (S and O) [45], mainly found in *Streptococcus* pathogenic species [45]; hemolysin II, III, X and β types [46] found in PFs mainly composed of z-proteobacteria [46,47] and firmicutes organisms [48]; shiga toxin, a common pathogenic factor in many virulent *E. coli* strains [49]; dermonecrotic toxin (DNT), one of the main virulence factors in many *Botretella* species [50] (pertussis in human), but at the same time present in plant pathogenic organisms like *Erwinia amylovora* [51] and *Erwinia pyrifoliae* [52]. The fact that we could find PFs containing DNT tagged as pathogens and others tagged as non-pathogenic (like the one containing DNT for *E. amylovora* and *E. pyrifoliae*) is an example of the ability of our clustering method to associate a given protein (a toxin in this case) to human pathogenicity as well as non-pathogenicity depending on the organism in which it is found.

Another example through which we could see the discriminative power of our PFs, was in associating pathogenicity to the different secretion system types proteins (SST1–SST6). For SST3 we identified 284 protein families, 147 of which were tagged as pathogenic. The pathogenic PFs were composed of human pathogenic z-proteobacteria strains, while the non-pathogenic PFs contained plant pathogenic organisms from proteobacteria genera like *Xanthomonas*, *Agrobacter* and *Erwinia*, which use SST3 (and other secretion systems) to infect the host cells of plants [53,54].

The protein families with high rank associated with non-pathogenicity (Table 6) were usually composed of proteins present in bacteria living in hot springs, lake surfaces or deep in the sea, and the functions are associated to their ability to survive under those extreme environmental conditions. Among those proteins are Ruberythrin, found in anaerobic sulphate-reducing bacteria like *Geobacter* and *Desulfitibrio* [55]. When the PFs were not composed of proteins from environmental bacteria, they contained mainly probiotics or plant pathogens. It is important to note that
since the WDM model was created with HT and LT parameters with values of 1.0 and 0 respectively, we only have PFs composed of proteins from either only pathogenic organisms or only non-pathogenic organisms.

Conclusions

There is an increasing need for fast identification of unknown bacteria with particular focus on the assessment of their potential pathogenicity. In this work we presented PathogenFinder, a web-server that by analysing the user-uploaded proteome can identify genomic features associated with both pathogenicity and non-pathogenicity. Given an input proteome the method quickly predicts its potential pathogenicity, making it a useful tool to be used together with other web-services developed for bacterial outbreak surveillance. Moreover, the possibility for the user to download the complete set of predicted pathogenicity features for the input organism makes PathogenFinder convenient for the analysis of pathogenic and harmless strains for microbiologists, epidemiologist and in general institutions studying bacterial pathogenesis.

One of the novel aspects in our approach is in the construction of the prediction models, which was carried out without any prior analysis of the proteins in the training-set, by just tagging our organisms as pathogenic or non-pathogenic and identifying protein families that were frequently found in pathogenic or non-pathogenic organisms.

It is important to notice that even though an isolate may have been obtained from a non-pathogenic environmental or animal or human related source it is not necessarily non-pathogenic. Such strain might in fact be highly pathogenic opportunistic pathogens. This naturally makes the creation of the optimal reference database difficult, but with increased number of isolates with well-defined meta data this is should still be doable.

We observed how PathogenFinder performs better than other pathogenicity prediction methods described in the literature, which usually rely on taxonomy and global sequence similarity with small sets of genes known to be associated with bacterial pathogenesis. We had less good results for species of the tenericutes phylum, and extra work need to be done to obtain statistically significant results for opportunistic strains (e.g. S. aureus) for which we could not tag any of our strains as non-pathogenic. The accuracy in predicting opportunistic bacteria could be improved by building specific models (e.g. at species level) as soon as new strains are available and there is a reasonable amount of both pathogenic and harmless strains. We have also shown how the prediction accuracy can be enhanced by increasing the number of organisms in the training-sets and/or making specific models at different taxonomic ranks, showing the example of E. coli, which is particularly difficult to predict because of the high similarity between commensal and pathogenic strains.

With the fast growing number of available bacterial complete genomes and with the increasing quality of the meta data we envision the possibility in the near future to build prediction models targeting only bacteria of a given genus or species, or even better, to build models to identify pathogenic features involved in specific diseases.

Materials and Methods

Training and Test Data

All available complete bacterial genomes (NCBI Genome Project, accessed on 10th Nov. 2010) were considered for the creation of the training-sets.

The pathogenicity information for the retrieved organisms were taken from NCBI genome project pages as described in Andreatta et al. [31], and for 885 of the 1,224 downloaded organisms, we were able to find pathogenicity information. The final complete training-set (Table S1) was composed of 513 organisms tagged as human non-pathogens and 372 tagged as human pathogens. For the human pathogenic organisms we checked for evidence in the literature. Opportunistic pathogens (e.g. from species like Staphylococcus aureus [56] or Pseudomonas aeruginosa [57]) were still tagged as pathogenic even though it has been shown that some of them can live inside the host without causing any disease, and their pathogenicity is sometimes related to the host’s health conditions.

From January 2012, NCBI removed pathogenicity information from its pages, redirecting the users to Genomes Online Database (GOLD) [58]. On 26th Feb. 2012 we queried GOLD for pathogenicity information about organisms that had been published after 5th Nov. 2010 (the date of the latest published bacteria in the training-set). We were able to extract pathogenicity information for 449 organisms, and subsequently retrieved the corresponding complete genomes and plasmids from NCBI based on the NCBI project ids.

The final test data (Table S1) was composed of 449 organisms, 294 of which were tagged as human non-pathogens and 155 as human pathogens.

Protein Clustering

The model creation consisted of the following 2 main steps:

I. Protein Clustering

II. PFDB Creation

The initial idea for clustering the proteins was to use BLAST [59], but due to the size of our dataset (almost 3 million proteins), it would not have been computationally feasible. Instead, we used CD-HIT [60], which made it possible to cluster all the proteins in approximately 24 days using 2× 3 Ghz dual-core CPUs in parallel and a 8 Gb of RAM.

The output from the program were 3 files containing respectively: 1) a list of cluster ids followed by the FASTA headers of the sequences composing the clusters; 2) a FASTA file containing all the clusters representative sequences; 3) a FASTA file containing all the solitary sequences that could not be included in any cluster.

Protein Family Database (PFDB) Creation

Our prediction models are based on the concept of protein families as initially proposed in Adreatta et al. [31]. Protein families are groups of proteins with a certain degree of similarity. The PFs were created using a two-steps filtering of the clusters created using CD-HIT. To perform this filtering we used four parameters: MinORG, Pratio, LT and HT.

Let ORG be the number of organisms which have proteins in a given cluster i. We define MinORG as the minimum number of organisms that must have proteins in the i cluster for it to be considered significant. As such, MinORG is a lower threshold for the ORG value.

Equation 1. Ratio of human pathogenic organisms having proteins inside the cluster i on the total number of organisms having proteins in i.

\[
P_{\text{ratio}}(i) = \frac{HP}{ORTH} \tag{1}
\]

\(P_{\text{ratio}}\) (Equation 1) is the ratio of the number of pathogen organisms having proteins in the i cluster (HP) on the total
number of organisms in \( i \) (ORG). LT and HT are thresholds for the \( P_{ratio} \) that we used to define if a given significant cluster should be tagged as pathogenic or non-pathogenic according to equation 2.

### Equation 2.

Function used to define if a given significant cluster should be tagged as 'pathogen family' or 'non-pathogen family'.

\[
f(i) = \begin{cases} 
-1 & \text{if } P_{ratio}(i) \leq LT \\
0 & \text{if } LT < P_{ratio}(i) < HT \\
1 & \text{if } P_{ratio}(i) \geq HT 
\end{cases}
\]  

\( (2) \)

Let \( f \) (Equation 2) be the function we use to decide if a given significant cluster should be tagged as pathogenic or non-pathogenic. If the number of sequences from pathogens and non-pathogens is too close in a given cluster (if \( P_{ratio} = 0.5 \) then \( f(i) = 0 \)), the cluster does not have any discriminative value for pathogenicity and is unusable.

Given a protein cluster \( i \), it was considered a protein family if the following 3 conditions were satisfied:

I. \( \text{ORG} \geq \text{MinORG} \)

II. \( f(i) \neq 0 \)

III. \( P_{ratio} \geq HT \) or \( P_{ratio} \leq LT \)

The significance of a protein family depends on its ORG value and its \( P_{ratio} \). A statistical measure called Z-score \( (Z) \) was used to take into account the above two values of a family and assess its significance. The estimation of the \( Z \) values was performed on the set \( C \) composed of all the clusters \( i \) satisfying condition I. Let \( \mu \) and \( \sigma \) be the average and standard deviation respectively of the \( P_{ratio} \) of the clusters in \( C \). \( Z \) is a measure representing by how many standard deviations \( \sigma \) the mean \( \mu \) of a sample (a cluster in our case) differs from the mean \( \mu \) of the population. Given a cluster \( i \) in \( C \), its mean corresponds to its \( P_{ratio} \), and we calculate the \( Z \) value for \( i \) as follows:

\[
Z_i = \frac{P_{ratio}(i) - \mu}{SE_i}
\]

Where \( SE \) is the standard error of the mean for \( i \), and it is:

\[
SE_i = \frac{\sigma}{\text{ORG}_i}
\]

To each protein family, a \( Z \) value was assigned, and these are used in the calculation of the final prediction score as well as a ranking value in the analysis of the protein families. Figure 1 shows the distributions of the \( P_{ratio} \) values and \( Z \)-scores for both significant clusters and protein families for the TM-Betaproteobacteria model, while Figure 2 shows for each of the models built the proportion of pathogenic and non-pathogenic families in the PFDB, together with the training-set and test-set for the 10 models built. All the sequences in the PFDB are used to perform the predictions.

### Models Optimisation

The prediction models were verified by 5-fold cross validation. For each of the models, many trials and tests were performed before choosing the MinOrg, LT and HT parameters for the final models. At each CV a parameter called \( Z_{thr} \) was further optimised. \( Z_{thr} \) is the threshold used to decide whether an input organism should be predicted as pathogenic or not, by comparing it to the summation of \( Z \) values obtained for the matching sequences in the input proteome. The parameters (MinOrg, HT, LT) (Table 1) of the models with the highest MCC in the CV tests were used to create the final models, and the corresponding \( Z_{thr} \) values will be used as thresholds for the predictions.

### Pathogenicity Prediction

The prediction method takes as input a FASTA file containing the proteins of the organism for which we want to assess the potential pathogenicity. In case the input is a complete or draft genome, initial gene prediction is performed using PRODIGAL [61]. PRODIGAL outputs a set of proteins representing the predicted genes. This is then used as input to our method. Using CD-HIT-2D [60], the input file is compared to the PFDB, and the output will contain all the input sequences that matched sequences in the PFDB, and that are used to compute the final prediction.

The following 4 steps describe the process that leads to the prediction:

I. Compare the input proteins to the PFDB

II. Filter hits based on the identity threshold (Equation 3)

III. Calculate final score summing the \( Z \) values associated to the matched PFs

IV. Compare the final score to the model’s \( Z_{thr} \) threshold and give the final prediction.

From the comparison in step I, we obtain a list of clusters, the representatives of which are sequences belonging to the PFDB, while the non-representative sequences come from the input. Because it is possible that more than one of the input proteins fall inside the same cluster, the sequence with the highest identity percentage with the representative is chosen. [hit].

#### Equation 3.

Calculates the identity threshold to select significant matches that will be used in the final prediction. The calculation is based on statistics on the identity values obtained for all matching query sequences.

\[
\text{id}en\text{Thr(hits)} = \begin{cases} 
\frac{\mu + \sigma + \text{max}}{2} & \text{if } \mu + \sigma \leq \text{max} \\
\text{max} & \text{otherwise}
\end{cases}
\]

The list of matches is then filtered based on an identity threshold that is dynamically computed at each prediction using the function idenThr (Equation 3). Let \( \text{hits} \) be a set containing all the percentage identity values for all our matches. Let \( \mu \) and \( \sigma \) be respectively the average and standard deviation of the percentage identity values in \( \text{hits} \). Let \( \text{max} \) be the maximum percentage identity obtained for the hits in PFDB. Remembering that, based on the settings of CD-HIT-2D, the minimum identity is 60%. Equation 3 calculates the identity threshold as the middle point between the maximum, and the average increased by one standard deviation, of the identities in \( \text{hits} \). Selecting all the matches with an identity higher than idenThr(hits), we will obtain a list of hits with a very high identity relatively to the distribution of identities of our hits.

The matches below that threshold will not be used in the final prediction. The process will sometimes greatly reduce the number of matches, but this is in favour of matches with higher identity, making the final prediction more reliable, if compared to the
results obtained using a fixed threshold, as we proved by using the paired student’s t-test (results not shown).

In the end we compute the summation of the Z-scores associated with the families matching the input sequences (III). If the sum of the Z-scores is above 2σ the input is considered pathogenic, otherwise it is considered non pathogenic (IV).

Supporting Information

Table S1  Training and Test organisms. xlsx file containing the list of organisms in the training and test-set and a table showing the phyla of the organisms in the training-set used to build the COMPL model.

(XLSX)

Table S2  Extra Escherichia Coli Strains. xlsx file containing the training-sets used for building ecoli_boost and enterobac_boost models, including the list of extra E. coli strains and a summary of the results in the prediction of E. coli and enterobacteraeae organisms.

(XLSX)

Table S3  Comparison with other methods. xlsx file containing a detailed description of the comparison of Pathogen-Finder and the method described in [28].

(XLSX)

Acknowledgments

We are grateful to Federico De Masi for his help in the phase of submission of the manuscript.

Author Contributions

Conceived and designed the experiments: SC OL. Performed the experiments: SC. Analyzed the data: SC MVL. Contributed reagents/materials/analysis tools: MVL FMA. Wrote the paper: SC. Helped review the manuscript: MVL FMA OL.

References