PathogenFinder

Johanne Ahrenfeldt
PhD Student
johah@cbs.dtu.dk
Who am I

Johanne Ahrenfeldt
johah@cbs.dtu.dk

• PhD student in Genomic Epidemiology
• Graduate engineer in Bioinformatics and Systems Biology from DTU – 2014
• Mainly work with Whole Genome based Phylogeny
Bacterial pathogenecity and virulence

• **Pathogenicity.** This is the potential capacity of certain species of microbes to cause an infectious process.

• **Virulence.** signifies the degree of pathogenicity of the given strain. Virulence, therefore, is an index of the qualitative individual nature of the pathogenic microorganism.
Microbes and humans

Very few microbes are always pathogenic
“Strict pathogens”

Many microbes are potentially pathogenic
“Opportunistic pathogens”

Most microbes are never pathogenic
“Non-pathogenic”
Student activation

• Give an example on a strict pathogen

• Give an example on an opportunistic pathogen

• Give an example on a non-pathogen
How do we know that a given pathogen causes a specific disease?

- **Koch's postulates**
  - the pathogen must be present in every case of the disease
  - the pathogen must be isolated from the diseased host & grown in pure culture
  - the specific disease must be reproduced when a pure culture of the pathogen is inoculated into a healthy susceptible host
  - the pathogen must be recoverable from the experimentally infected host
Use 2 minutes to discuss in small groups how you would conquer the island.

Include:

• **How to get on to and how to stay on the island**
  *Back-paddle, throw an anchor, use a rope, swim from the boat (might require more than one swimmer!!)*

• **How to avoid being detected by the island defense**
  *Camouflage, hide, dig-in, costume*

• **How to eliminate the island defense**
  *Poison, weapon, scare to perform suicide*
Coordinated attack
Gene regulation – A tool for a coordinated attack

(a) An operon

P  O  Z  Y  A
Purpose

The main purpose of PathogenFinder is to predict the pathogenecity of a given bacteria, based on the whole genome sequence or the proteome.
Method

• PathogenFinder identifies and divides the genes after protein families

• The genes are the clustered using CD-hit

• After clustering it is determined whether a group of genes is more pathogenic or non-pathogenic
Pathogenic gene families

- **Org P1**: A - C - E - G
- **Org P2**: C - E - G
- **Org P3**: B - C - E - G
- **Org Pn**: B - C - E - G

- **Org N1**: B - C - D - E - H
- **Org N2**: B - C - D - E - H
- **Org N3**: B - C - D - F - H
- **Org Nn**: B - C - D - H
Pathogenic gene families

Table 1. 10 top scoring pathogenicity families, and function of their members.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Z-score</th>
<th>P</th>
<th>N</th>
<th>Function of proteins in the family</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.29</td>
<td>42</td>
<td>4</td>
<td>Mutarotases, YjhT proteins</td>
</tr>
<tr>
<td>2</td>
<td>8.25</td>
<td>33</td>
<td>1</td>
<td>Fimbrial proteins, putative adhesins</td>
</tr>
<tr>
<td>3</td>
<td>8.12</td>
<td>38</td>
<td>3</td>
<td>Proteins of unknown function</td>
</tr>
<tr>
<td>4</td>
<td>8.02</td>
<td>40</td>
<td>4</td>
<td>Cytochrome b&lt;sub&gt;562&lt;/sub&gt;</td>
</tr>
<tr>
<td>5</td>
<td>7.89</td>
<td>39</td>
<td>4</td>
<td>Proteins of unknown function</td>
</tr>
<tr>
<td>6</td>
<td>7.86</td>
<td>36</td>
<td>3</td>
<td>Methyltransferases</td>
</tr>
<tr>
<td>7</td>
<td>7.82</td>
<td>30</td>
<td>1</td>
<td>Fimbrial proteins, pilin proteins</td>
</tr>
<tr>
<td>8</td>
<td>7.56</td>
<td>25</td>
<td>0</td>
<td>Heat shock proteins, DNA-repair</td>
</tr>
<tr>
<td>9</td>
<td>7.46</td>
<td>36</td>
<td>4</td>
<td>5-carboxymethyl-2-hydroxymuconate isomerase</td>
</tr>
<tr>
<td>10</td>
<td>7.06</td>
<td>25</td>
<td>1</td>
<td>Type III secretion proteins, path. island proteins</td>
</tr>
</tbody>
</table>
Predicting pathogenicity

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

I    Compare the input proteins to the PathogenFinder Database of protein families

II   Filter hits based on the identity threshold

III  Calculate final score summing the Z values associated to the matched PFs

IV   Compare the final score to the model’s Zthr threshold and give the final prediction
PathogenFinder 1.1

View the version history of this server.

Choose the phylum or class of your organism:
Choose 'All' if you want to use the model created using all bacteria

Automatic Model Selection

Sequencing Platform
Select the sequencing platform used to generate the uploaded reads. (Note: Select 'Assembled Genome' if you are uploading preassembled reads)

Proteome

Isolate File

Name | Size | Progress | Status
--- | --- | --- | ---

Upload | Remove

CITATIONS
For publication of results, please cite:

- PMID: 24204795 doi: 10.1371/journal.pone.0077302
## Results

The input organism was predicted as human pathogen

<table>
<thead>
<tr>
<th>Probability of being a human pathogen</th>
<th>0.888</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input proteome coverage (%)</td>
<td>6.42</td>
</tr>
<tr>
<td>Matched Pathogenic Families</td>
<td>308</td>
</tr>
<tr>
<td>Matched Not Pathogenic Families</td>
<td>17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sequences</th>
<th>5062</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total bpp</td>
<td>1608055</td>
</tr>
<tr>
<td>Longest seq</td>
<td>3164</td>
</tr>
<tr>
<td>Shortest seq</td>
<td>30</td>
</tr>
<tr>
<td>Avg seq length</td>
<td>317.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>NODE_157_length_219841_cov_31.485369_86 # 101449 # 103950 # 1 # ID=118_86;partial=00;start_type=ATG;rbs_motif=GGAGG;rbs_spacer=5-10bp;gc_cont=0.604</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matched Family</td>
<td>21069 AP006725  Klebsiella pneumoniae NTUH-K2044 DNA, complete genome. Gammaproteobacteria phosphoenolpyruvate-protein phosphotransferase</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NODE_159_length_245710_cov_33.035236_14 # 16609 # 19041 # 1 # ID=120_14;partial=00;start_type=ATG;rbs_motif=GGAG/GAG;rbs_spacer=5-10bp;gc_cont=0.599</td>
</tr>
<tr>
<td>Matched Family</td>
<td>21069 AP006725  Klebsiella pneumoniae NTUH-K2044 DNA, complete genome. Gammaproteobacteria putative formate acetyltransferase 3</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NODE_14_length_236341_cov_29.808062_145 # 140782 # 142776 # -1 # ID=14_145;partial=00;start_type=ATG;rbs_motif=GGA/GAG/AGG;rbs_spacer=5-10bp;gc_cont=0.566</td>
</tr>
<tr>
<td>Matched Family</td>
<td>21069 AP006725  Klebsiella pneumoniae NTUH-K2044 DNA, complete genome. Gammaproteobacteria phosphoglycerate transport system sensor protein</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NODE_65_length_274784_cov_33.074543_169 # 180284 # 182086 # -1 #</td>
</tr>
</tbody>
</table>
