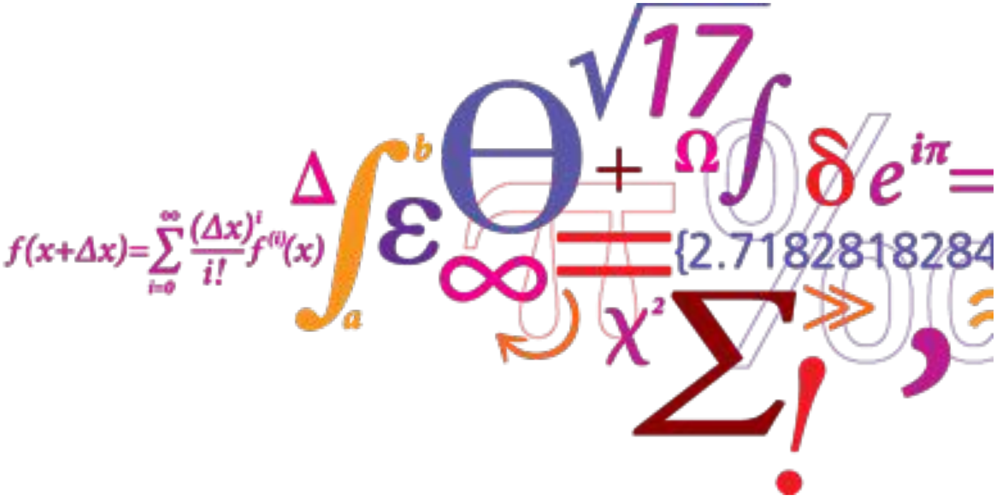
PathogenFinder

Johanne Ahrenfeldt
PhD Student
ja@bioinformatics.dtu.dk



DTU Bioinformatics
Department of Bio and Health Informatics

Who am I

Johanne Ahrenfeldt

ja@bioinformatics.dtu.dk

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

Today

- Pathogenicity
- PathogenFinder
- *Exercises*
- Virulence
- VirulenceFinder

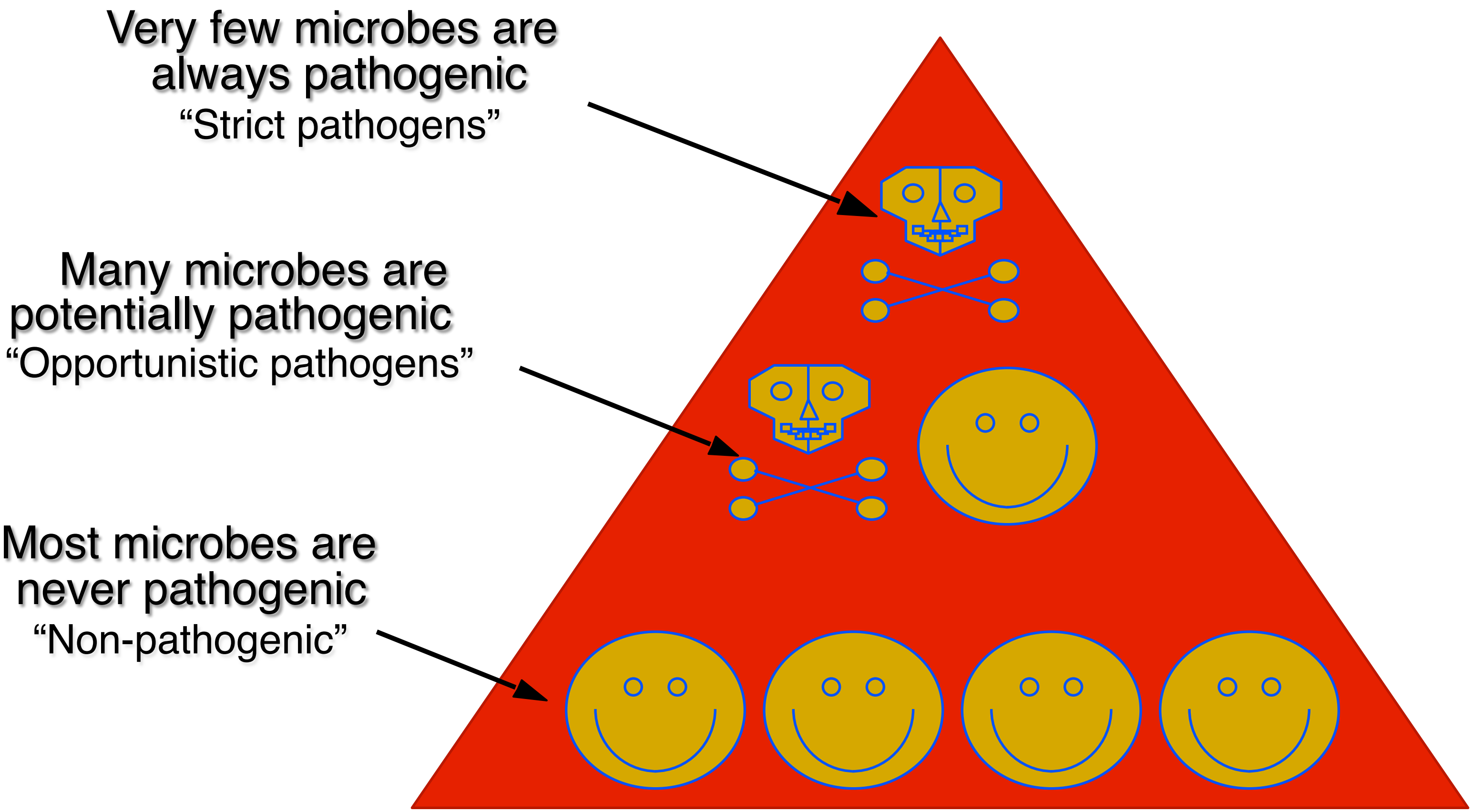
Lunch

- Antimicrobial resistance
- ResFinder
- *Exercises*

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

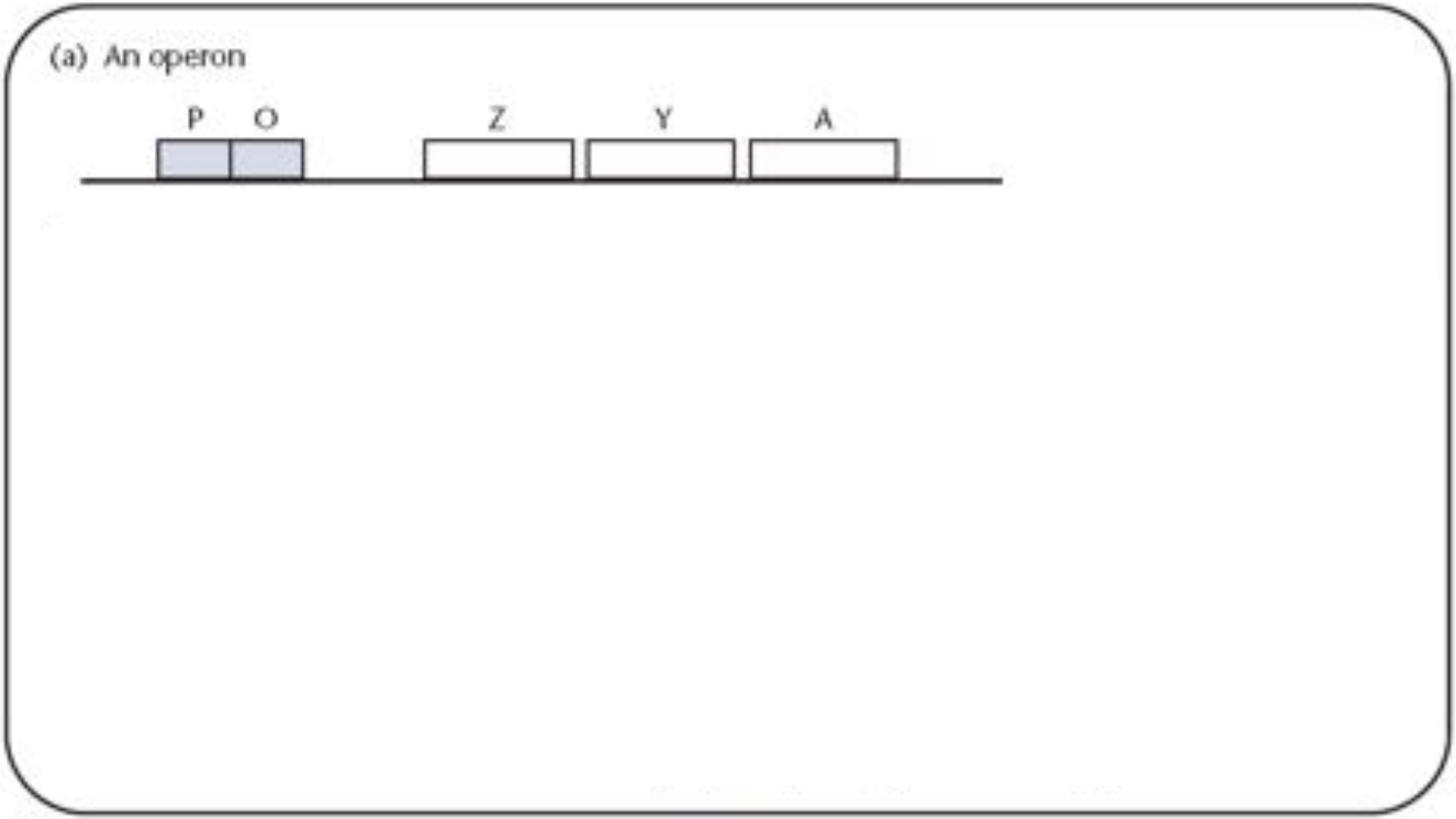
- **How to eliminate the island defense**

Poison, weapon, scare to perform suicide

Coordinated attack

GETTYSBURG
LEE
Cemetery Hill
Culp's Hill
MEADE
LONGSTREET
LITTLE ROUND TOP
BIG ROUND TOP
TWO TAVERNS

Gene regulation – A tool for a coordinated attack



PathogenFinder

OPEN ACCESS Freely available online

PLOS | ONE

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

Salvatore Cosentino[1*], Mette Voldby Larsen[1], Frank Møller Aarestrup[2], Ole Lund[1]

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.

Citation: Cosentino S, Voldby Larsen M, Møller Aarestrup F, Lund O (2013) PathogenFinder - Distinguishing Friend from Foe Using Bacterial Whole Genome Sequence Data. PLoS ONE 8(10): e77302. doi:10.1371/journal.pone.0077302

Editor: Jason G. Bertaux, University of Hawaii Manoa, United States of America

Received August 2, 2013; **Accepted** September 6, 2013; **Published** October 28, 2013

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

Purpose

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

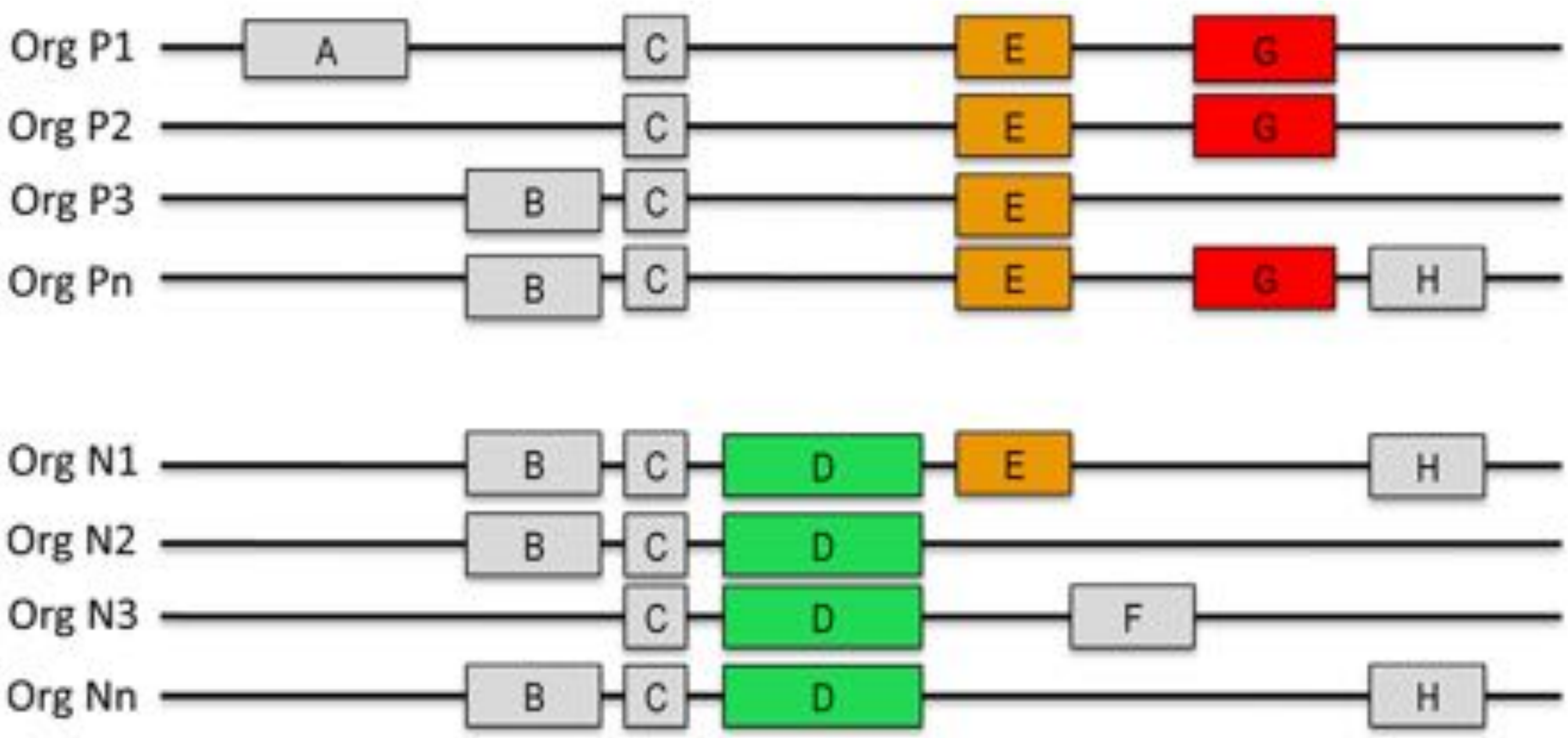
Method

- 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



Pathogenic gene families

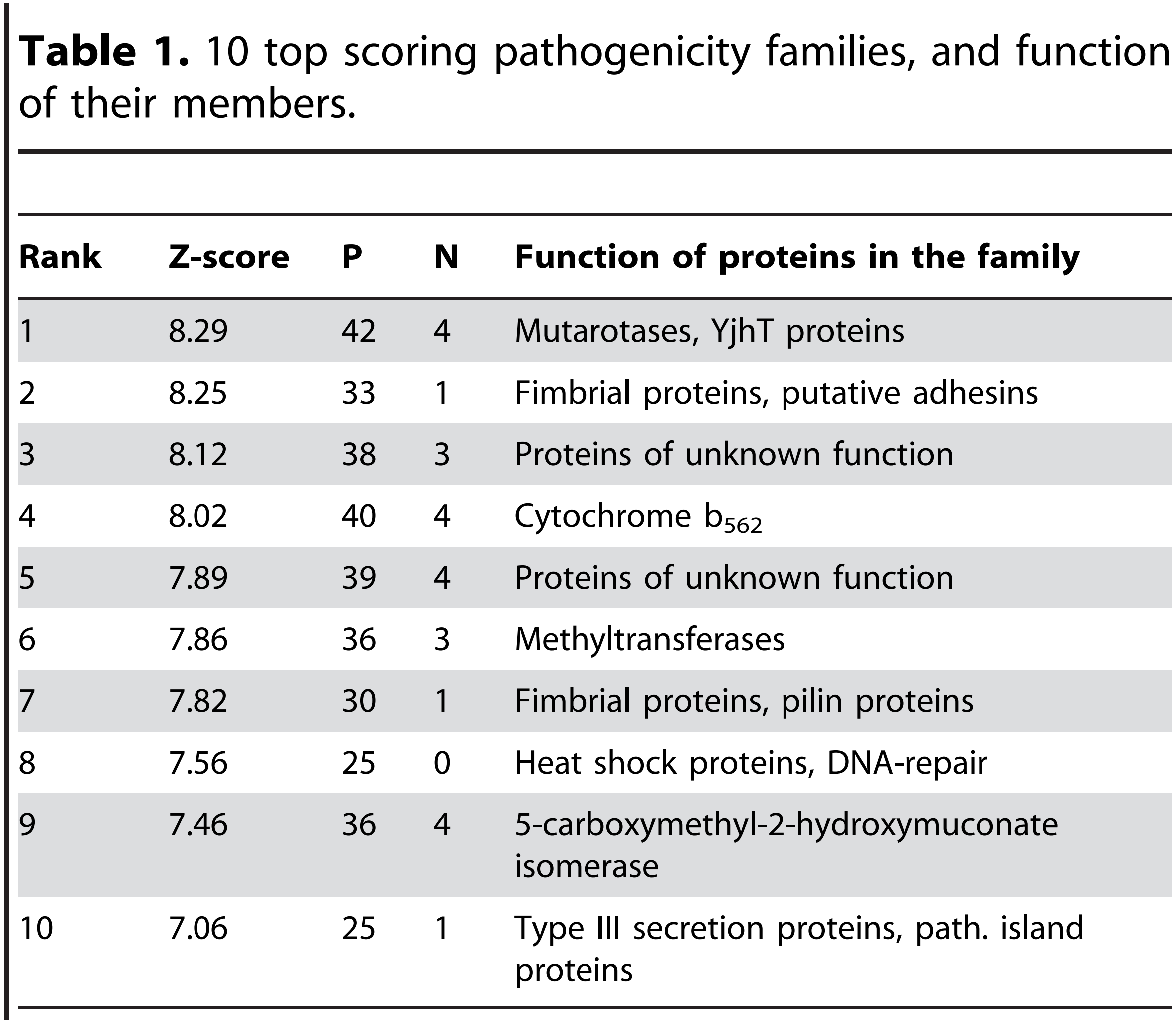


Table 1. 10 top scoring pathogenicity families, and function of their members.

Rank	Z-score	P	N	Function of proteins in the family
1	8.29	42	4	Mutarotases, YjhT proteins
2	8.25	33	1	Fimbrial proteins, putative adhesins
3	8.12	38	3	Proteins of unknown function
4	8.02	40	4	Cytochrome b_{562}
5	7.89	39	4	Proteins of unknown function
6	7.86	36	3	Methyltransferases
7	7.82	30	1	Fimbrial proteins, pilin proteins
8	7.56	25	0	Heat shock proteins, DNA-repair
9	7.46	36	4	5-carboxymethyl-2-hydroxymuconate isomerase
10	7.06	25	1	Type III secretion proteins, path. island proteins

Predicting pathogenicity

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

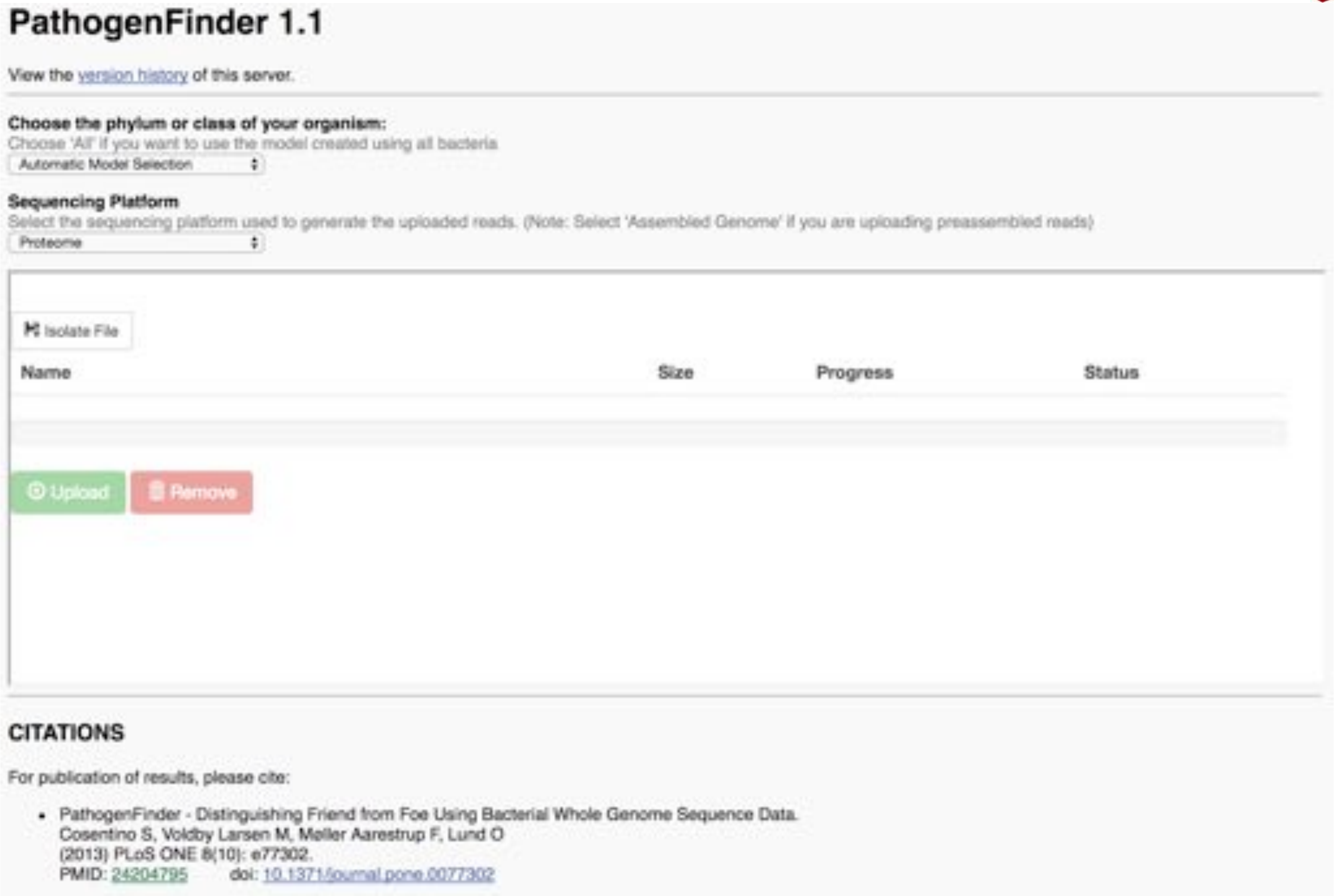
I Compare the input proteins to the PathogenFinder Database of protein families

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 Z-score threshold and give the final prediction

https://cge.cbs.dtu.dk/services/PathogenFinder/



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)

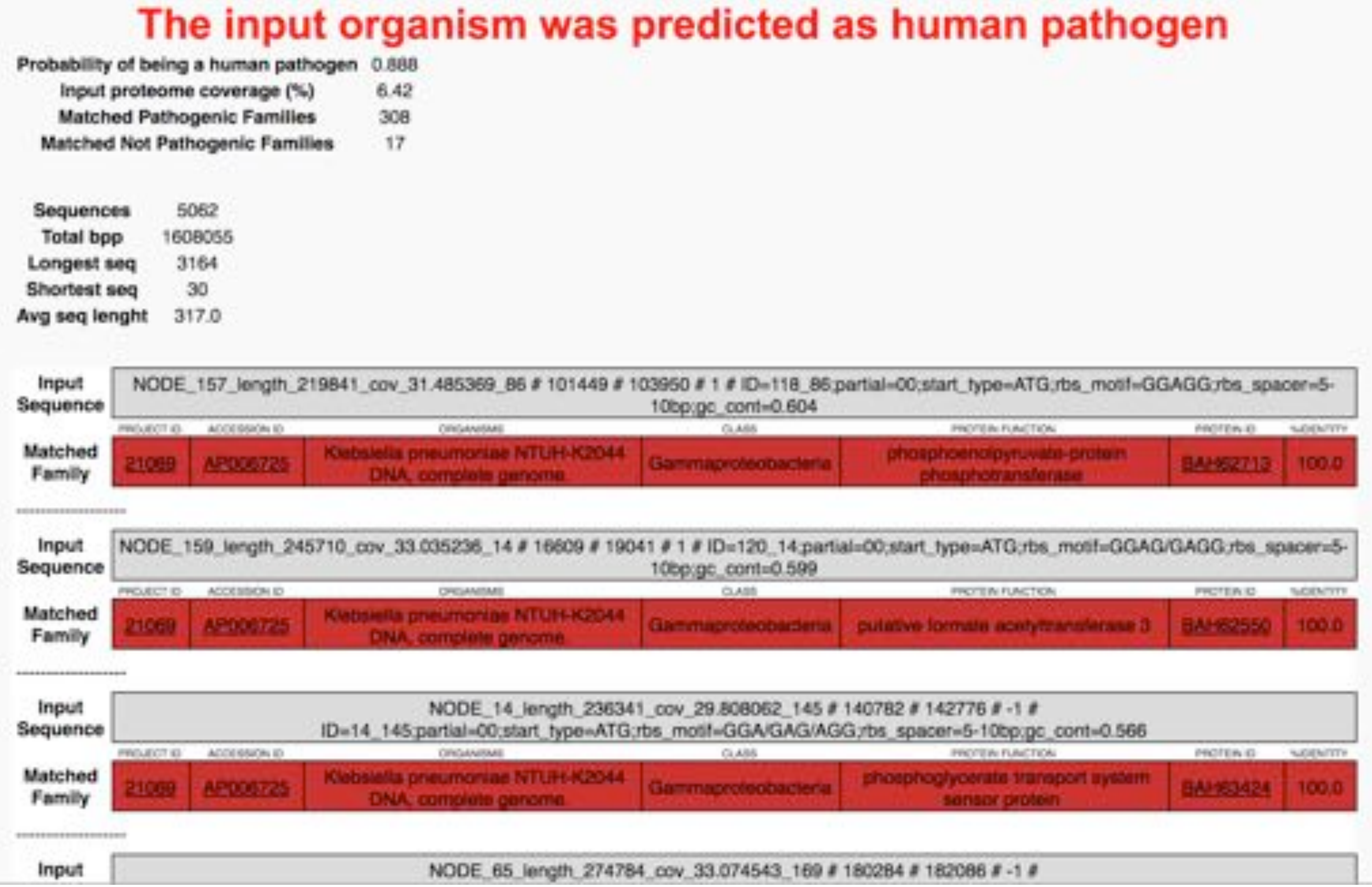
Proteome

CITATIONS

For publication of results, please cite:

- PathogenFinder - Distinguishing Friend from Foe Using Bacterial Whole Genome Sequence Data.
  Cosentino S, Voldby Larsen M, Møller Aarestrup F, Lund O
  (2013) PLoS ONE 8(10): e77302.
  PMID: 24204795 doi: 10.1371/journal.pone.0077302

Results



The input organism was predicted as human pathogen

Probability of being a human pathogen	0.888
Input proteome coverage (%)	6.42
Matched Pathogenic Families	308
Matched Not Pathogenic Families	17

Sequences	5062
Total bpp	1608055
Longest seq	3164
Shortest seq	30
Avg seq lenght	317.0

Input Sequence: 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

	PROJECT ID	ACCESSION ID	ORGANISMS	CLASS	PROTEIN FUNCTION	PROTEIN ID	%IDENTITY
Matched Family	21069	AP006725	Klebsiella pneumoniae NTUH-K2044 DNA, complete genome	Gammaproteobacteria	phosphoenolpyruvate-protein phosphotransferase	BAH62713	100.0

Input Sequence: NODE_159_length_245710_cov_33.035236_14 # 16609 # 19041 # 1 # ID=120_14;partial=00;start_type=ATG;rbs_motif=GGAG/GAGG;rbs_spacer=5-10bp;gc_cont=0.599

	PROJECT ID	ACCESSION ID	ORGANISMS	CLASS	PROTEIN FUNCTION	PROTEIN ID	%IDENTITY
Matched Family	21069	AP006725	Klebsiella pneumoniae NTUH-K2044 DNA, complete genome	Gammaproteobacteria	putative formate acetyltransferase 3	BAH62550	100.0

Input Sequence: 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

	PROJECT ID	ACCESSION ID	ORGANISMS	CLASS	PROTEIN FUNCTION	PROTEIN ID	%IDENTITY
Matched Family	21069	AP006725	Klebsiella pneumoniae NTUH-K2044 DNA, complete genome	Gammaproteobacteria	phosphoglycerate transport system sensor protein	BAH63424	100.0

Input: NODE_65_length_274784_cov_33.074543_169 # 180284 # 182086 # -1 #