BIOINFORMATICS APPLICATIONS NOTE Vol. 31 no. 4 2015, pages 590–592
doi:10.1093/bioinformatics/btu681

Genome analysis Advance Access publication October 17, 2014

mimicMe: a web server for prediction and analysis of host-like proteins in microbial pathogens

Pavel Petrenko and Andrew C. Doxey*
Department of Biology, University of Waterloo, Waterloo, Ontario, Canada

Associate Editor: Burkhard Rost

ABSTRACT

Summary: mimicMe is a web server for prediction and analysis of host-like proteins (*mimics*) encoded by microbial pathogens. Users select a host species and any set of pathogen and control proteomes (bacterial, fungal, protozoan or viral) and mimicMe reports host-like proteins that are unique to or enriched among pathogens. Additional server features include visualization of structural similarities between pathogen and host proteins as well as function-enrichment analysis.

Availability and implementation: mimicMe is available at http://mimicme.uwaterloo.ca

Contact: acdoxey@uwaterloo.ca

Received on June 30, 2014; revised on September 4, 2014; accepted on October 14, 2014

1 INTRODUCTION

Many pathogens encode virulence factors that mimic the function of one or more host proteins. These host-like proteins, called *mimics*, exploit host pathways and/or facilitate evasion of host immune detection (Bhavsar *et al*., 2007; Stebbins and Galan, 2001). Mimics may be homologous to their host counterparts, and originate through host-to-pathogen horizontal transfer, or may evolve through convergent evolution (Elde and Malik, 2009). Several studies have focused on the latter type, identifying pathogen mimics of short host peptide fragments (Hagai *et al*., 2014; Ludin *et al*., 2011). Here, we focus on pathogen mimicry involving host-pathogen sequence homology, which can be more easily detected through standard bioinformatic approaches.

In previous work, we used a BLAST-based, comparative proteomics approach to detect mimics of this type in human pathogenic bacteria (Doxey and McConkey, 2013). By detecting human-like proteins highly unique to pathogens and relatively absent in non-pathogens (controls), known mimics [e.g. *Legionella* RalF mimicry of human ADP-ribosylation factor guanine nucleotide exchange factors (ARF-GEFs)], and novel ones (e.g. multi-species mimics of human collagen and leucine-rich repeats) could be identified.

Here, we report mimicMe, a web server for exploration of host-like proteins and potential mimics in pathogens. mimicMe automates and extends our previous analysis to include a range of common host species from mammals to plants, and pathogens from bacteria, viruses, protozoans and fungi. The tool is best suited for predicting mimics that are homologous to host proteins, but other types of mimicry may be detected through altered parameters. The tool allows for structural visualization of predicted mimicry relationships as well as function-enrichment analysis to identify host gene families, functions or pathways that are targets of molecular mimicry by pathogens of interest. We anticipate that the tool will be useful in comparative analyses of pathogens, where it may serve to generate hypotheses for future experimental studies.

2 METHODS AND IMPLEMENTATION

2.1 Data sources

Proteome sequence data for selected hosts and microbes were retrieved from the NCBI's ftp resource (ftp://ncbi.nlm.nih.gov/genomes/) and the NCBI BioProject Database (Barrett *et al*., 2012). This included several host species of interest (e.g. *Arabidopsis*, human, cow, zebrafish) and 2765 bacterial, 2321 viral, 35 fungal and 49 protozoan proteomes.

2.2 Host versus microbial proteome BLASTing

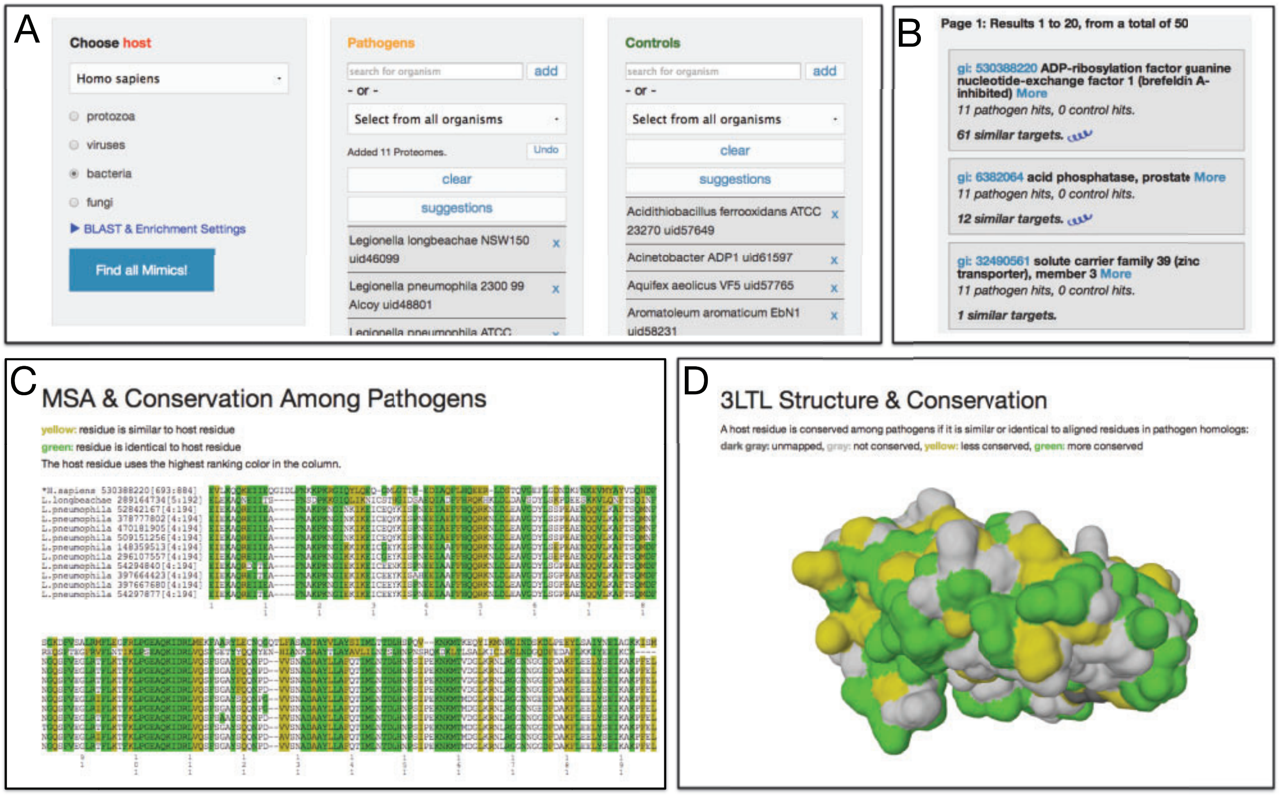
For mimicMe to rapidly predict molecular mimicry relationships in any selection of organisms, proteome-to-proteome alignments were precomputed for all possible host–microbe relationships. Computations were performed using the blast+ suite (Camacho *et al*., 2009) distributed among cluster nodes in SHARCNET, a high-performance computing network. After alignment output was parsed, all host and microbial protein pairs sharing at least one high scoring pair alignment with an *E*-value ≤ 0.05 were stored in a MongoDB database, along with corresponding alignment data.

2.3 mimicMe input/output

2.3.1 Queries The front end of mimicMe provides a simple web interface for selecting a host organism, and a set of pathogen species (where mimics are expected to be found) and control species (where mimics are not expected). Suggested lists of pathogens and controls are also provided as an option. The interface includes additional parameters to adjust the stringency of predictions, including an *E*-value threshold, as well as the minimum and maximum number of allowed hits in pathogens and non-pathogens, respectively [see (Doxey and McConkey, 2013) for more details]. All input parameters are automatically saved as workflows that can be bookmarked in the browser. Common workflows are are included in a section of mimicMe.

2.3.2 Results The output of a mimicMe analysis is a list of mimicry target proteins and associated mimics in the input set

*To whom correspondence should be addressed.


A
Choose host
Homo sapiens
○ protozoa
○ viruses
● bacteria
○ fungi
▶ BLAST & Enrichment Settings
Find all Mimics!
Pathogens
search for organism
add
- or -
Select from all organisms
Added 11 Proteomes.
Undo
clear
suggestions
Legionella longbeachae NSW150 uid46099 x
Legionella pneumophila 2300 99 Alcoy uid48801 x
Legionella pneumophila ATCC
Controls
search for organism
add
- or -
Select from all organisms
clear
suggestions
Acidithiobacillus ferrooxidans ATCC 23270 uid57649 x
Acinetobacter ADP1 uid61597 x
Aquifex aeolicus VF5 uid57765 x
Aromatoleum aromaticum EbN1 uid58231 x
B
C
MSA & Conservation Among Pathogens
*H.sapiens 530388220[693:884]
L.longbeachae 289164734[5:192]
L.pneumophila 52842167[4:194]
L.pneumophila 378777802[4:194]
L.pneumophila 470181905[4:194]
L.pneumophila 509151256[4:194]
L.pneumophila 148359513[4:194]
L.pneumophila 296107557[4:194]
L.pneumophila 54294840[4:194]
L.pneumophila 397664423[4:194]
L.pneumophila 397667680[4:194]
L.pneumophila 54297877[4:194]
D
3LTL Structure & Conservation

Page 1: Results 1 to 20, from a total of 50

gi: 530388220 ADP-ribosylation factor guanine nucleotide-exchange factor 1 (brefeldin A-inhibited) More
11 pathogen hits, 0 control hits.
61 similar targets.

gi: 6382064 acid phosphatase, prostate More
11 pathogen hits, 0 control hits.
12 similar targets.

gi: 32490561 solute carrier family 39 (zinc transporter), member 3 More
11 pathogen hits, 0 control hits.
1 similar targets.

yellow: residue is similar to host residue
green: residue is identical to host residue
The host residue uses the highest ranking color in the column.

A host residue is conserved among pathogens if it is similar or identical to aligned residues in pathogen homologs:
dark gray: unmapped, gray: not conserved, yellow: less conserved, green: more conserved

Fig. 1. Screenshots of the mimicMe resource. (**A**) Mimics are identified based on user selection of host, pathogen and control proteomes. BLAST and enrichment parameters are also adjustable. Results page (**B**), sequence alignment (**C**) and structural visualization (**D**) of an example involving predicted mimicry of human ARF-GEFs by *Legionella* RalF proteins. (Color version of this figure is available at *Bioinformatics* online.)

of pathogens that meet the specified criteria. Results are sorted by pathogen versus non-pathogen abundance and alignment score, searchable by keyword to focus on functions of interest, and similar (redundant) predictions are grouped together to reduce redundancy. PSORTb (Yu *et al.*, 2010) predictions of subcellular localization are also available for most bacterial proteins.

3 SAMPLE ANALYSIS AND MIMICME FEATURES

As an example analysis, suppose one aims to identify mimics conserved in multiple species of *Legionella* that are absent in broad range of non-pathogens (Fig. 1A). mimicMe predicts a ranked list of 50 mimicry relationships (Fig. 1B) unique to *Legionella*. The top prediction involves detected mimicry of the Sec7 domain of human ARF-GEFs by the *Legionella* RalF protein, a known mimic and virulence factor, present in 11 *Legionella* proteomes and 0 controls (Fig. 1B). Further analysis can be performed using some of mimicMe's built-in tools described below.

3.1 Multiple sequence alignment

mimicMe offers the user the option of performing a MUSCLE (http://www.drive5.com/muscle/) multiple sequence alignment (MSA) between a host protein and all of its pathogen hits (Fig. 1C). The alignment is colored based on sequence similarities to the host protein, to highlight particular regions or motifs mimicked by the pathogen/s.

3.2 Structural visualization

The conservation pattern from the MSA can also be mapped onto a known PDB structure (http://www.pdb.org) of the host protein, and visualized using GLmol (http://webglmol.sourceforge.jp) (Fig. 1D). This may highlight particular regions of structural mimicry.

3.3 Function enrichment

The user is also provided with the option to determine statistically enriched functions among the set of predicted mimicry targets (host proteins), which may reflect common host functions targeted by the pathogen. Gene enrichment analysis is performed using goatools (https://github.com/tanghaibao/goatools), with Gene Ontology annotations retrieved from the EBI Gene Ontology Annotation project (Camon *et al.*, 2004). In the example described above, mimicMe returns GO:0016192 ("vesicle-mediated transport") as the top enriched function (P = 1.18e–5). Indeed, manipulation of host vesicular trafficking is a hallmark of *Legionella* pathogenesis (Ge and Shao, 2011).

4 CONCLUSION

mimicMe provides an automated pipeline for BLAST-based detection and exploration of host-like proteins (mimics) in microbial pathogens. Predictions made by mimicMe may serve as a guide for experimentalists interested in mimicry-related mechanisms of host–microbe interactions and virulence.

Funding: This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), through a Discovery Grant to A.C.D.

Conflict of interest: none declared.

REFERENCES

Barrett,T. *et al*. (2012) BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res*., **40.D1**, D57–D63.

Bhavsar,A.P. *et al*. (2007) Manipulation of host-cell pathways by bacterial pathogens. *Nature*, **449**, 827–834.

Camacho,C. *et al*. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Camon,E. *et al*. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res*., **32**, D262–D66.

Doxey,A.C. and McConkey,B.J. (2013) Prediction of molecular mimicry candidates in human pathogenic bacteria. *Virulence*, **4**, 453–466.

Elde,N.C. and Malik,H.S. (2009) The evolutionary conundrum of pathogen mimicry. *Nat Rev Microbiol*., **7**, 787–797.

Ge,J. and Shao,F. (2011) Manipulation of host vesicular trafficking and innate immune defence by Legionella Dot/Icm effectors. *Cell Microbiol*., **13**, 1870–1880.

Hagai,T. *et al*. (2014) Use of host-like peptide motifs in viral proteins is a prevalent strategy in host-virus interactions. *Cell Rep*., **7**, 1–11.

Ludin,P. *et al*. (2011) Genome-wide identification of molecular mimicry candidates in parasites. *PLoS One*, **6**, e17546.

Stebbins,C.E. and Gálan,J.E. (2001) Structural mimicry in bacterial virulence. *Nature*, **412**, 701–705.

Yu,N.Y. *et al*. (2010) PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, **26**, 1608–1615.