

mimicDetector: a pipeline for protein motif mimicry detection in host-pathogen interactions

Kaylee D. Rich^{1,*},  and James D. Wasmuth^{1,*}, 

¹Faculty of Veterinary Medicine, University of Calgary, Calgary, Alberta, T2N 4Z6, Canada

*Corresponding authors. Kaylee D. Rich, Faculty of Veterinary Medicine, University of Calgary, 3280 Hospital Drive NW, Calgary, Alberta, T2N 4Z6, Canada. E-mail: kaylee.rich@ucalgary.ca and James D. Wasmuth, Faculty of Veterinary Medicine, University of Calgary, 3280 Hospital Drive NW, Calgary, Alberta, T2N 4Z6, Canada. E-mail: jwasmuth@ucalgary.ca

Associate Editor: Jianlin Cheng

Abstract

Motivation: Molecular mimicry is used by pathogens to evade the host immune system and manipulate other host cellular processes. It is often mediated by short motifs in non-homologous proteins, whose detection challenges the sensitivity and specificity of existing bioinformatics tools.

Results: We present mimicDetector, a *k*-mer-based pipeline for identifying protein-level molecular mimicry between pathogens and their hosts. Applied to 17 globally important pathogens, mimicDetector identified a broad and biologically plausible set of mimicry candidates, including helminth proteins mimicking components of the human complement system and a *Leishmania infantum* mimic of Reticulon-4, a regulator of immune cell recruitment.

Availability and implementation: mimicDetector is freely available at <https://github.com/kayleerich/mimicDetector/>, implemented in Python and Snakemake, and compatible with Unix-based systems.

1 Introduction

Pathogens often evade immune detection through molecular mimicry, where their molecules structurally or functionally imitate those of the host. Mimicry has been described in glycans, lipids, and nucleic acids, with protein mimicry arguably the most widely studied (van Die and Cummings 2010, Buck *et al.* 2014, Bosch *et al.* 2021, Maguire *et al.* 2024). Identifying mimicry is important for understanding immune evasion, identifying therapeutic targets and revealing co-evolution.

Protein mimicry typically involves short motifs (5-to-20 amino acids in length), that mediate host-pathogen interactions by altering signaling, protein binding, or immune recognition [as reviewed by Sámano-Sánchez and Gibson (2020)]. However, these short motifs are difficult to detect computationally. Current state-of-the-art motif discovery tools, including those from the MEME suite or SLiMfinder, are designed to find overrepresented motifs among non-homologous, functionally related sequences (e.g. co-regulated genes) (Edwards *et al.* 2007, Bailey *et al.* 2015). Other tools, such as SLiMPred and SLiMDisc, predict motifs documented in curated databases like ELM (Mooney *et al.* 2012, Palopoli *et al.* 2015). These approaches work well for short linear motifs (SLiM) and post-translational modification (PTM) sites, but assume recurrence for prior motif knowledge. These assumptions do not hold

for molecular mimicry, where motifs are typically rare, non-recurrent, and largely unknown. Consequently, many mimicry exploration studies rely on local-local sequence alignment tools despite their default parameters being poorly suited for short sequence fragments (Ludin *et al.* 2011, Doxey and McConkey 2013, Pearson 2013, Armijos-Jaramillo *et al.* 2021, Emiliani *et al.* 2022).

We previously resurrected a *k*-mer-based strategy that aligned pathogen-derived peptides against host and control proteomes (Ludin *et al.* 2011, Rich *et al.* 2023). While effective, it relied on strict identity thresholds that may exclude biologically relevant mimics with subtle sequence divergence. Additionally, homologue filtering—intended to reduce noise—risked discarding genuine mimics.

Here, we present mimicDetector, an optimized and scalable pipeline for identifying mimicry motifs across divergent proteomes without prior motif knowledge. Unlike existing motif-finding tools, mimicDetector addresses a distinct biological problem with different biological assumptions: detecting short, non-recurrent mimicry motifs that other algorithms are not designed to uncover. To achieve this, we systematically evaluated alignment algorithms, scoring matrices, homologue handling, and filtering strategies to improve sensitivity and specificity.

Received: 22 May 2025. Revised: 27 November 2025. Accepted: 11 December 2025

© The Author(s) 2026. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

2 Benchmarking and optimization

2.1 Overall approach

To guide tool selection and parameter optimization for mimicDetector, we designed benchmarking experiments using identical-hit and similar-hit peptide datasets from three well-annotated, phylogenetically diverse reference proteomes used in prior mimicry studies: *Plasmodium falciparum* 3D7, *Mycobacterium tuberculosis* H37Rv, and *Brugia malayi* (Ludin et al. 2011, Muthye and Wasmuth 2023, Rich et al. 2023, Yanik et al. 2023). A full list of the proteomes used in this study is provided in Table S1, available as [supplementary data](#) at *Bioinformatics* online.

We evaluated the recall performance of five protein alignment search tools (BLASTP, DIAMOND, PHMMER, TOPAZ, and Glam2Scan), comparing substitution matrices (BLOSUM62 and PAM30), and, where possible, the *wordsize* (2 and 3) used for seeding the alignments (Fig. S1, available as [supplementary data](#) at *Bioinformatics* online) (Altschul et al. 1990, Frith et al. 2008, Eddy 2011, Buchfink et al. 2015, Medlar and Holm 2018). Peptides were extracted as overlapping *k*-mers (lengths 5-to-14), and 10 sets of 1000 *k*-mers were randomly sampled from the proteome (Li 2018). We used each tool and various parameter combinations to align these *k*-mers back to each original proteome (Table S2, available as [supplementary data](#) at *Bioinformatics* online), and used a custom script to measure the number of identical and similar *k*-mers, allowing up to *k*-3 mismatches.

2.2 Alignment recall

BLASTP consistently outperformed other alignment tools for both identical and similar alignments (Figs S2 and S3, available as [supplementary data](#) at *Bioinformatics* online). For exact matches, the PAM30 substitution matrix outperformed BLOSUM62 when $k \leq 9$ and was equivalent at $k \geq 10$. Reducing *wordsize* to 2 improved recall for $k=5$, where seeding was a limiting factor. For similar-hit alignments, BLASTP again led, though performance varied. PAM30 outperformed BLOSUM62 when: (i) $k \leq 9$; (ii) $k \geq 10$ with few mismatches. In contrast, BLOSUM62 was slightly better PAM30 for the longer *k*-mers. Reducing *wordsize* helped with PAM30-scored alignments, but had little effect on BLOSUM62-scored alignments.

TOPAZ performed well for identical *k*-mer alignments. However, it was excluded due to a mismatch-counting error and weaker similar-hit performance. DIAMOND and PHMMER underperformed compared to BLASTP across all tests. Glam2Scan showed perfect recall for identical-hits but was excluded due to excessive runtimes (Fig. S4, available as [supplementary data](#) at *Bioinformatics* online).

2.3 Proteome coverage

Beyond recall, we assessed how much of the proteome was matched by the query *k*-mers. We consider it desirable to find a higher coverage of the target proteome, so reducing the likelihood of excluding potentially relevant motifs early in the pipeline. Merging overlapping alignments, we found that BLASTP with PAM30 and *wordsize*=2 provided the highest coverage,

plateauing at $k=12$ (Fig. S5, available as [supplementary data](#) at *Bioinformatics* online).

2.4 False discovery rate

We estimated the empirical false discovery rate (FDR) using a target-decoy approach, with the human proteome as the target and shuffled human sequences as the decoy:

$$\text{FDR} = c \cdot (N_{\text{decoy}}/N_{\text{target}}) \quad (1)$$

where N_{target} is the number of hits to the target database, N_{decoy} is the number of hits to the decoy database, and c is a correction factor for the number of unique peptides (UP) in the target and decoy databases (Käll et al. 2008, Lee et al. 2021):

$$c = 1/(UP_{\text{decoy}}/UP_{\text{target}}) \quad (2)$$

We calculated FDR for runs of mimicDetector for combinations of threshold value ranges for bitscore difference ($0 \leq b \leq 3$), minimum E-value (range $0.05 \geq e \geq 0.0005$) and minimum solvent accessibility ($q=0.50$ or 0.75) for the three aforementioned species. We found that combinations of $0 \leq b \leq 2$, $0.01 \geq e \geq 0.001$, and $q=0.75$ produced $\text{FDR} < 0.05$ for all three species but runs with $b=1$ and $e=0.001$ consistently had the lowest FDR, regardless of dataset (Fig. S6, available as [supplementary data](#) at *Bioinformatics* online).

2.5 Homologue removal

The original pipeline filtered full-length pathogen proteins with significant alignments to host proteins, presumably to reduce the number of *k*-mers to consider in later searches. However, this may exclude proteins due to a shared promiscuous domain or biologically relevant mimics contains within homologous genes that duplicated and undergone subfunctionalization (Pastrana et al. 1998, Armijos-Jaramillo et al. 2021, Huang et al. 2022). We found that removing this step led to a small increase in the number of candidate motifs in *P. falciparum* and a marked increase in helminth species, which are discussed in the next section (Fig. S7, available as [supplementary data](#) at *Bioinformatics* online).

3 Implementation

mimicDetector integrates optimized alignment parameters and refined filtering strategies into a streamlined, modular pipeline (Fig. 1). We selected the BLASTP search algorithm, with the PAM30 substitution matrix and *wordsize*=2 (Table S3, available as [supplementary data](#) at *Bioinformatics* online). The default *k*-mer length is 12 amino acids. A further change to early versions is the use of bitscore, rather than percent identity, to compare *k*-mer alignments.

Users provide proteomes for the pathogen, host, and negative control species (typically non-pathogen) in FASTA format. The pathogen proteome is fragmented into overlapping *k*-mers, which serve as queries for ungapped local-local alignments.

Post-alignment, mimicDetector applies a series of biologically motivated filters. To achieve an $\text{FDR} < 0.05$ (see section 2.4), we

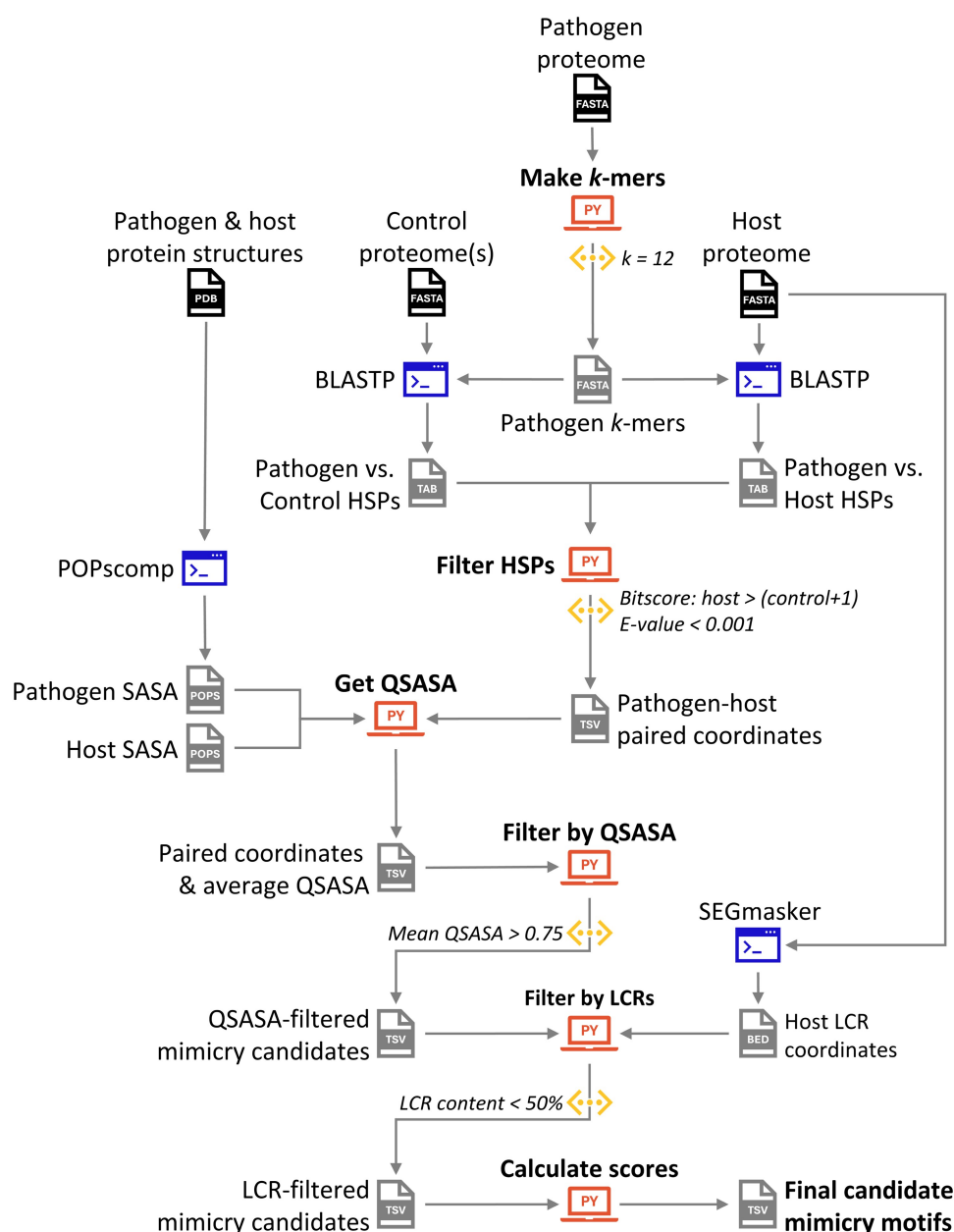


Figure 1 Workflow for the mimicDetector pipeline with suggested thresholds.

retain alignments where E-value is ≤ 0.001 and the pathogen-host alignment bitscore exceeds that of the best pathogen-control alignment by at least one bit. Overlapping k -mer alignments are merged into a single preliminary mimicry candidate. Finally, hits with a mean solvent accessible surface area $>75\%$, via POPSCOMP, are kept, and hits where $>50\%$ of the sequence is low-complexity, via Segmasker, are removed (Wootton and Federhen 1996, Kleinjung and Fraternali 2005).

The final output includes a ranked list of predicted mimicry candidates, which includes pathogen and host protein sequences with positional information and recalculated alignment scores and E-values. The pipeline and its accompanying documentation are freely available on GitHub.

4 Results and discussion

We applied mimicDetector to 17 eukaryotic pathogen proteomes to identify protein-level mimicry targeting the human proteome (Table S1, available as supplementary data at Bioinformatics online). The updated pipeline consistently identified more mimicry candidates than our previous approach, particularly when homologous proteins were retained and evaluated at the k -mer level.

In *Leishmania infantum*, we identified two notable mimics. The first protein, LINF_180015000 (UniProt: A4HXX0), shares a motif with CD244 (2B4; UniProt: Q9BZW8), an immune checkpoint receptor expressed on natural killer cells (De Freitas et al.

2021, Sanz *et al.* 2021). The second protein, LINF_310020500 (UniProt: A4I6R4), shared a motif with Reticulon-4B (Nogo-B; UniProt: Q9NQC3), a modulator of leukocyte migration and TLR9 trafficking, which has been implicated in Leishmania-driven changes to macrophage behavior (Kimura *et al.* 2015, Bhattacharya *et al.* 2023).

In helminths (parasitic worms), mimicDetector uncovered possible mimicry of complement components. Four *Schistosoma mansoni* proteins had similarity to C1qA and four nematode species matched C1qB (Table S4, available as supplementary data at Bioinformatics online). Many helminth species, including those mentioned here, secrete a calreticulin homologue which inhibits classical activation of the complement pathway through interaction with complement protein C1q (Floudas *et al.* 2017, Yadav *et al.* 2021, Esperante *et al.* 2023). Functional assays are planned to examine the potential interactions suggested by mimicDetector.

Finally, we note that the search algorithms tested here were not originally designed for our specific goal. For example, the long runtimes for Glam2Scan were in part due to formatting the data appropriately for each search. Future mimicry detection may benefit from using underlying algorithms, such as Waterman-Eggert (Waterman and Eggert 1987).

Author contributions

Kaylee Rich (Conceptualization [equal], Funding acquisition [supporting], Methodology [lead], Software [lead], Validation [lead], Writing—original draft [lead], Writing—review & editing [supporting]), James Wasmuth (Conceptualization [equal], Funding acquisition [lead], Project administration [lead], Supervision [lead], Writing—original draft [supporting], Writing—review & editing [lead])

Supplementary data

Supplementary data is available at Bioinformatics online.

Conflict of interest

None declared.

Funding

This work was supported by the Natural Science and Engineering Research Council of Canada (NSERC) [Discovery grant #04589-2020 to J.D.W.] and the University of Calgary [Eyes High doctoral scholarship to K.D.R.].

Data availability

No new data were generated or analysed in support of this research. The data analysed for this article are available from Uniprot using the accession codes in Table S1 of the supplementary material.

References

- Altschul SF, Gish W, Miller W *et al.* Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10.
- Armijos-Jaramillo V, Espinosa N, Vizcaíno K *et al.* A novel In silico method for molecular mimicry detection finds a formin with the potential to manipulate the maize cell cytoskeleton. *Mol Plant Microbe Interact* 2021;**34**:815–25.
- Bailey TL, Johnson J, Grant CE *et al.* The MEME suite. *Nucleic Acids Res* 2015;**43**:W39–49.
- Bhattacharya P, Gannavaram S, Ismail N *et al.* Toll-like receptor-9 (TLR-9) signaling is crucial for inducing protective immunity following immunization with genetically modified live attenuated leishmania parasites. *Pathogens* 2023;**12**:534.
- Bosch M, Sweet MJ, Parton RG *et al.* Lipid droplets and the host–pathogen dynamic: FATal attraction? *J Cell Biol* 2021;**220**:e202104005.
- Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;**12**:59–60.
- Buck AH, Coakley G, Simbari F *et al.* Exosomes secreted by nematode parasites transfer small RNAs to mammalian cells and modulate innate immunity. *Nat Commun* 2014;**5**:5488.
- De Freitas E, Silva R, Von Stebut E. Unraveling the role of immune checkpoints in leishmaniasis. *Front Immunol* 2021;**12**:620144.
- van Die I, Cummings RD. Glycan gimmickry by parasitic helminths: a strategy for modulating the host immune response? *Glycobiology* 2010;**20**:2–12.
- Doxey AC, McConkey BJ. Prediction of molecular mimicry candidates in human pathogenic bacteria. *Virulence* 2013;**4**:453–66.
- Eddy SR. Accelerated profile HMM searches. Pearson WR (ed.). *PLoS Comput Biol* 2011;**7**:e1002195.
- Edwards RJ, Davey NE, Shields DC. SLIMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. *PLoS One* 2007;**2**:e967.
- Emiliani Y, Muzi G, Sánchez A *et al.* Prediction of molecular mimicry between proteins from Trypanosoma sp. and human antigens associated with systemic lupus erythematosus. *Microb Pathog* 2022;**172**:105760.
- Esperante D, Flisser A, Mendlovic F. The many faces of parasite calreticulin. *Front Immunol* 2023;**14**:1101390.
- Floudas A, Cluxton CD, Fahel J *et al.* Composition of the Schistosoma mansoni worm secretome: identification of immune modulatory Cyclophilin A. *PLoS Negl Trop Dis* 2017;**11**:e0006012.
- Frith MC, Saunders NFW, Kobe B *et al.* Discovering sequence motifs with arbitrary insertions and deletions. Stormo G (ed.). *PLoS Comput Biol* 2008;**4**:e1000071.
- Huang S, Qiu Y, Ma Z *et al.* A secreted MIF homologue from Trichinella spiralis binds to and interacts with host monocytes. *Acta Trop* 2022;**234**:106615.
- Käll L, Storey JD, MacCoss MJ *et al.* Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J Proteome Res* 2008;**7**:29–34.
- Kimura T, Endo S, Inui M *et al.* Endoplasmic protein Nogo-B (RTN4-B) interacts with GRAMD4 and regulates TLR9-mediated innate immune responses. *J Immunol* 2015;**194**:5426–36.

- Kleinjung J, Fraternali F. POPSCOMP: an automated interaction analysis of biomolecular complexes. *Nucleic Acids Res* 2005; **33**:W342–6.
- Lee S, Park H, Kim H. Comparison of false-discovery rates of various decoy databases. *Proteome Sci* 2021; **19**:11.
- Li H. seqtk. 2018. <https://github.com/lh3/seqtk>
- Ludin P, Nilsson D, Mäser P. Genome-Wide identification of molecular mimicry candidates in parasites. El-Sayed N (ed.). *PLoS ONE* 2011; **6**:e17546.
- Maguire C, Wang C, Ramasamy A *et al.* Molecular mimicry as a mechanism of viral immune evasion and autoimmunity. *Nat Commun* 2024; **15**:9403.
- Medlar A, Holm L. TOPAZ: asymmetric suffix array neighbourhood search for massive protein databases. *BMC Bioinformatics* 2018; **19**:278.
- Mooney C, Pollastri G, Shields DC *et al.* Prediction of short linear protein binding regions. *J Mol Biol* 2012; **415**:193–204.
- Muthye V, Wasmuth JD. Proteome-wide comparison of tertiary protein structures reveals molecular mimicry in plasmodium-human interactions. *Front Parasitol* 2023; **2**:1162697. <https://doi.org/10.3389/fpara.2023.1162697>
- Palopoli N, Lythgow KT, Edwards RJ. QSLiMfinder: improved short linear motif prediction using specific query protein data. *Bioinformatics* 2015; **31**:2284–93.
- Pastrana DV, Raghavan N, FitzGerald P *et al.* Filarial nematode parasites secrete a homologue of the human cytokine macrophage migration inhibitory factor. *Infect Immun* 1998; **66**:5955–63.
- Pearson WR. Selecting the right Similarity-Scoring matrix. *Curr Protoc Bioinf* 2013; **43**:3.5.1–.5.9.
- Rich KD, Srivastava S, Muthye VR *et al.* Identification of potential molecular mimicry in pathogen-host interactions. *PeerJ* 2023; **11**: e16339.
- Sámano-Sánchez H, Gibson TJ. Mimicry of short linear motifs by bacterial pathogens: a drugging opportunity. *Trends Biochem Sci* 2020; **45**:526–44.
- Sanz CR, Miró G, Sevane N *et al.* Modulation of host immune response during Leishmania infantum natural infection: a Whole-Transcriptome analysis of the popliteal lymph nodes in dogs. *Front Immunol* 2021; **12**:794627.
- Waterman MS, Eggert M. A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *J Mol Biol* 1987; **197**:723–8.
- Wootton JC, Federhen S. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol* 1996; **266**:554–71.
- Yadav S, Sharma P, Sharma A *et al.* Immunization with *Brugia malayi* calreticulin protein generates robust antiparasitic immunity and offers protection during experimental lymphatic filariasis. *ACS Infect Dis* 2021; **7**:790–9.
- Yanik S, Venkatesh V, Parker ML *et al.* Structure guided mimicry of an essential *P. falciparum* receptor-ligand complex enhances cross neutralizing antibodies. *Nat Commun* 2023; **14**:5879.