University of Manitoba

Public Health Agency of Canada
Agence de la santé publique du Canada

MEDICAL MICROBIOLOGY
INFECTIOUS DISEASES

CAVES: A Novel Tool for Comparative Analysis of Variant Epitope Sequences

Katherine Li [1, 2], Connor Lowey [3], Paul Sandstrom [1, 2], Hezhao Ji [1, 2]

[1] National Microbiology Laboratories at JC Wilt Infectious Diseases Research Centre, Public Health Agency of Canada

[2] Department of Medical Microbiology and Infectious Diseases, University of Manitoba

[3] Independent programmer

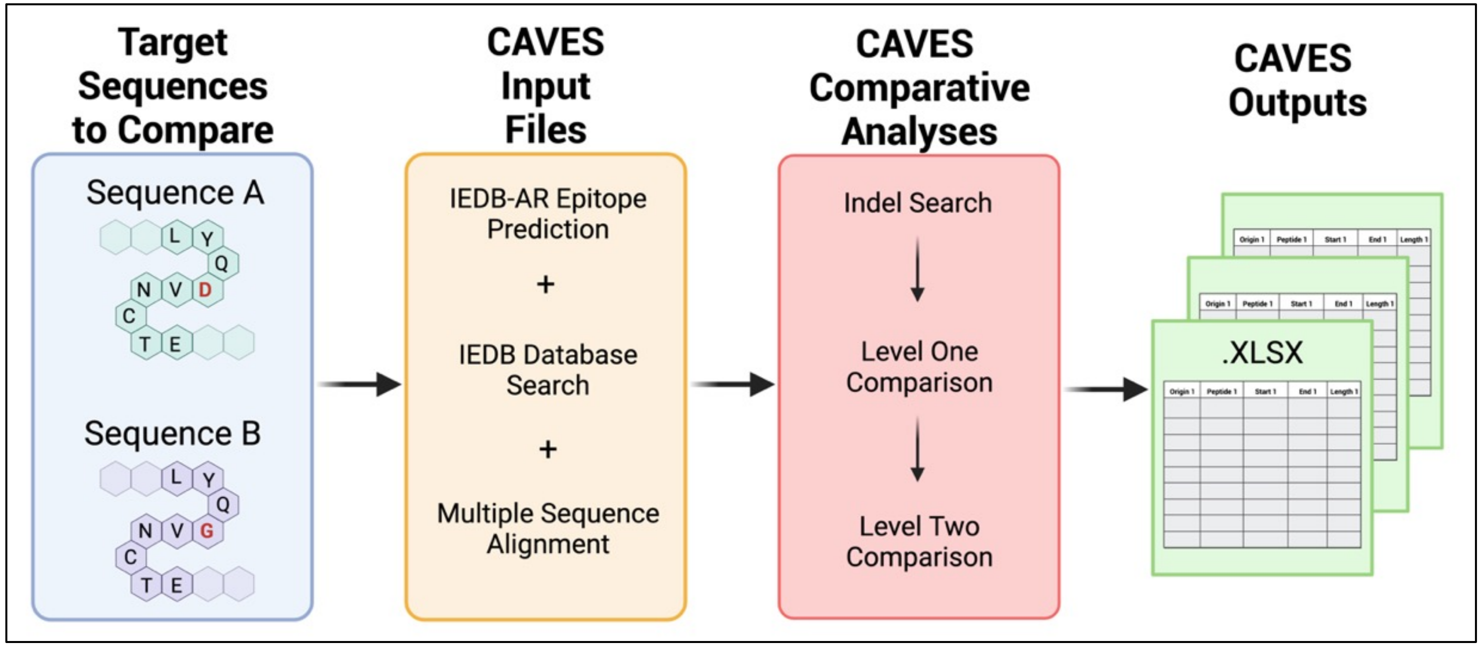
Correspondence: lik3456@myumanitoba.ca

The authors have no conflict of interest to declare

Introduction

- Epitopes are part of an antigen that is recognizable by the host immune system and elicits a specific immune response
- Understanding how epitope recognition differs between pathogens is important for vaccine and therapeutic design
- Putative epitopes can be predicted using computational-based epitope analysis programs such as the IEDB-AR
- Manual comparison of massive lists of epitope sequences from different pathogen strains is laborious, time-consuming, and prone to human error, often making it unfeasible

IEDB - Immune Epitope Database; IEDB-AR – IEDB-Analysis Resource



Comparative Analysis of Variant Epitope Sequences (CAVES)

- A novel tool developed for automated comparative analyses of epitopes from two closely related pathogens (*Sequence A* vs *Sequence B*)
- Takes epitope data from the IEDB as input, and outputs results in .XLSX format (Microsoft Excel)
- Uses two comparison levels to determine the similarities/differences between epitopes from the compared sequences and their relevance in published literature
- Runs through a graphical user interface on Windows operating systems and is freely available at https://github.com/connor-lowey/CAVES

Matching Criteria

- *CAVES compares epitope sequences (as amino acid peptides) between two given pathogens (Sequence A vs B)*
- *Sorts each epitope into a category based on the degree to which it matches with epitopes from the opposing sequence*

Exact Match

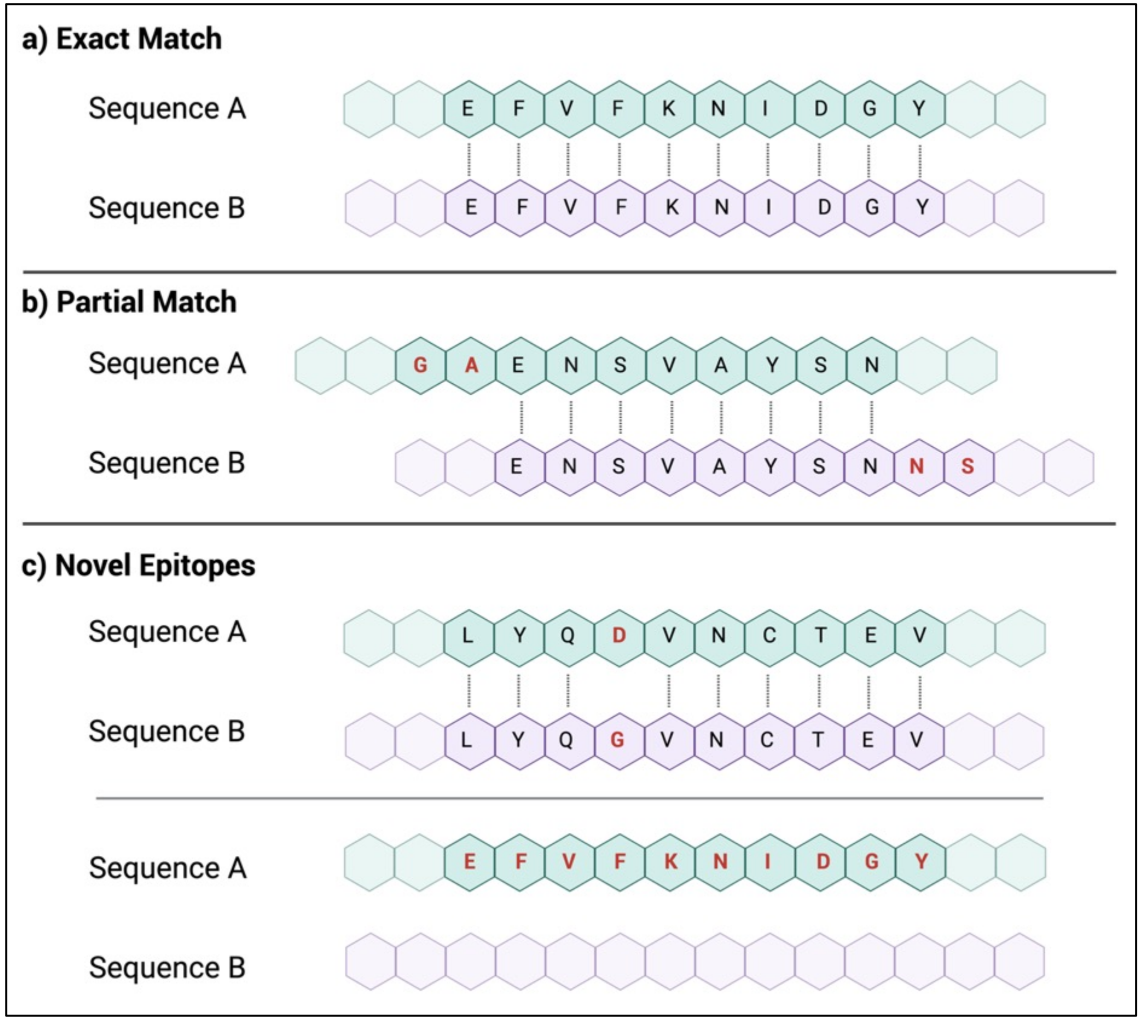
- When two epitopes have identical amino acid characters at the same sequence loci
- Must match for the entire length of at least one of the two epitopes being compared

Partial Match

- When two epitopes have identical amino acid characters at the same sequence loci but are offset from each other
- Offset sequences means the match cannot possibly cover the entire length of either epitope

Novel Epitopes

- When two epitopes create a match of any length (Exact or Partial) but contain a mutation (substitution, insertion, or deletion), making them distinctly unique epitopes
- **Or**, when an epitope did not find a match (of any length) with the opposing sequence

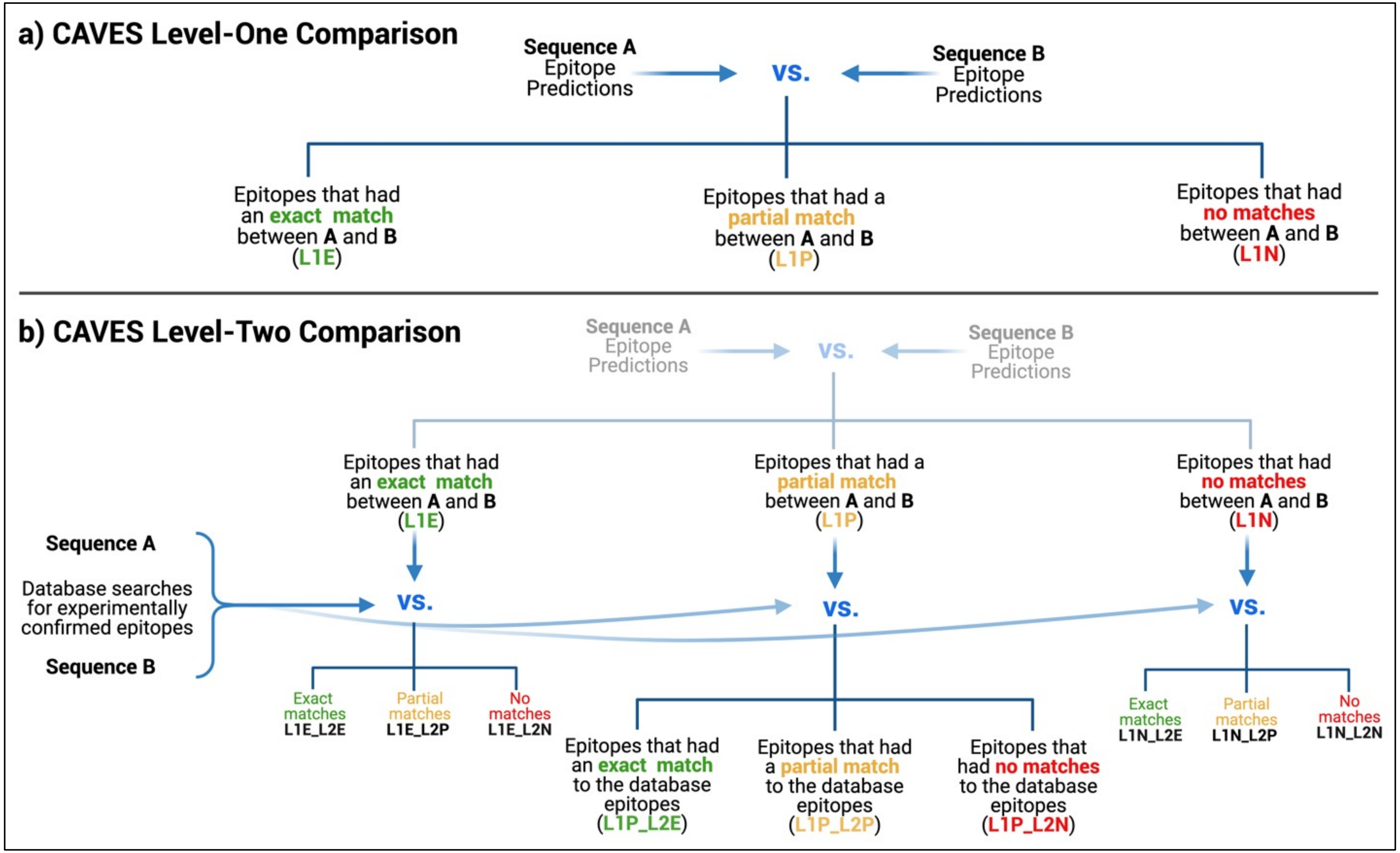


Two-Level Approach

- *Each comparison level sorts epitopes into categories of Exact matches, Partial matches, or Novel epitopes*

CAVES Level-One compares epitope predictions between the two pathogens (*Sequence A* vs *B*) to determine their similarities and differences

CAVES Level-Two compares epitopes from each sorted list (generated in Level-One) against epitopes from a database query to determine which epitope predictions have been experimentally confirmed in published literature



a) CAVES Level-One Comparison
Sequence A Epitope Predictions
vs.
Sequence B Epitope Predictions
Epitopes that had an exact match between A and B (L1E)
Epitopes that had a partial match between A and B (L1P)
Epitopes that had no matches between A and B (L1N)
b) CAVES Level-Two Comparison
Sequence A Epitope Predictions
vs.
Sequence B Epitope Predictions
Epitopes that had an exact match between A and B (L1E)
Epitopes that had a partial match between A and B (L1P)
Epitopes that had no matches between A and B (L1N)
Sequence A
Database searches for experimentally confirmed epitopes
Sequence B
vs.
vs.
vs.
Exact matches L1E_L2E
Partial matches L1E_L2P
No matches L1E_L2N
Epitopes that had an exact match to the database epitopes (L1P_L2E)
Epitopes that had a partial match to the database epitopes (L1P_L2P)
Epitopes that had no matches to the database epitopes (L1P_L2N)
Exact matches L1N_L2E
Partial matches L1N_L2P
No matches L1N_L2N

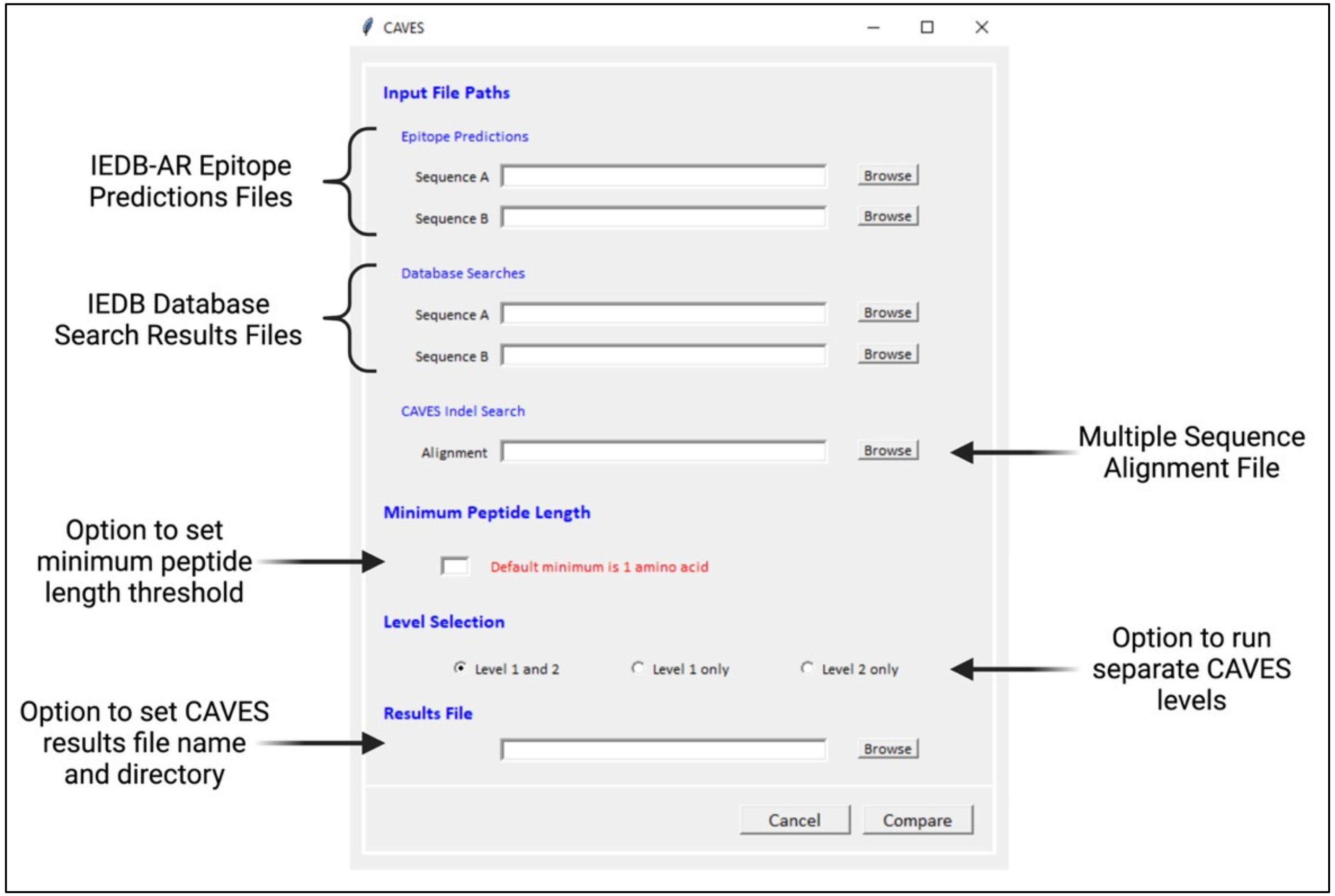
Test Dataset

Two SARS-CoV-2 spike protein sequences
(*Wuhan strain* vs. *Alpha VOC strain*)

- T cell HLA II epitopes predicted for each sequence using the IEDB-AR TepiTool
- The IEDB database of experimentally confirmed epitopes queried for each sequence

Results:

- CAVES accurately binned all epitopes into the Exact, Partial, and Novel categories for Level-One and Two
- CAVES Novel categories correctly identified all epitopes covering characteristic Alpha VOC mutations



● Level 1 and 2 ○ Level 1 only ○ Level 2 only

IEDB-AR Epitope Predictions Files

IEDB Database Search Results Files

Multiple Sequence Alignment File

Option to set minimum peptide length threshold

Option to run separate CAVES levels

Option to set CAVES results file name and directory

Conclusion

- CAVES greatly reduces time and user workload
  - Compared and sorted test dataset (1,129 total epitopes) in 3.6 seconds
- Highly applicable for the study of any hypermutable pathogen such as HIV-1
- Can be used for evolutionary analyses or to compare epitopes from different prediction tools for computational validation

VOC - Variant of Concern; IEDB - Immune Epitope Database; IEDB-AR – IEDB-Analysis Resource

Research Manitoba

University of Manitoba

MEDICAL MICROBIOLOGY INFECTIOUS DISEASES

Public Health Agency of Canada

Agence de la santé publique du Canada