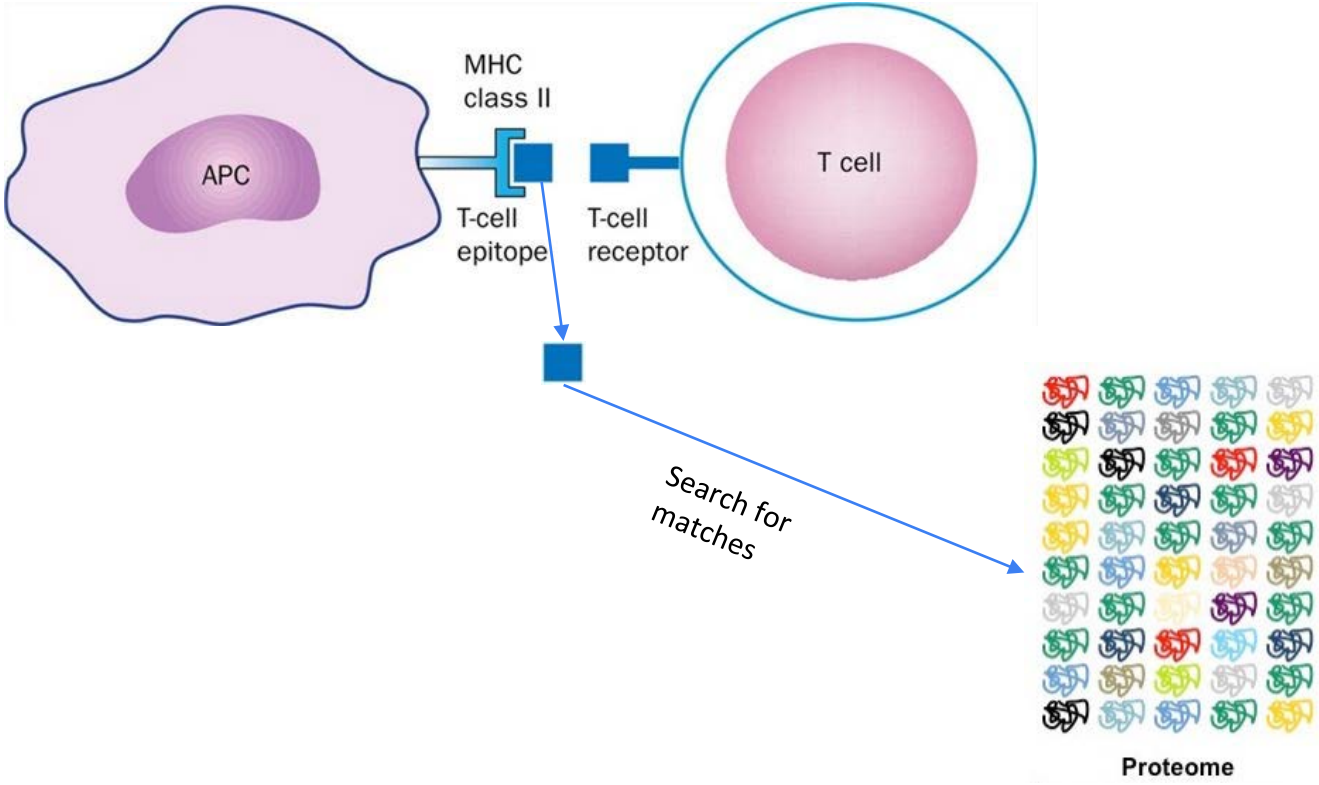
PEPMatch: A Tool to Identify Short Peptide Sequence Matches in Large Sets of Proteins

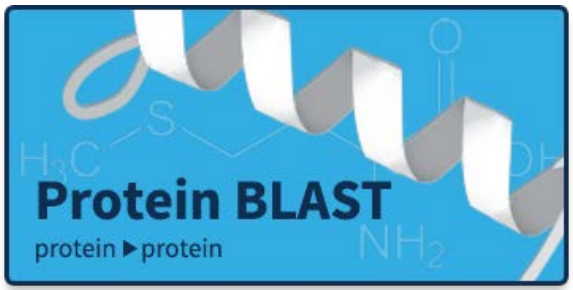
Presented by: Daniel Marrama, Associate Bioinformatics Specialist I

Peptide Sequence Matching



* Set of all proteins expressed by an organism

Common Tools for Peptide Sequence Matching

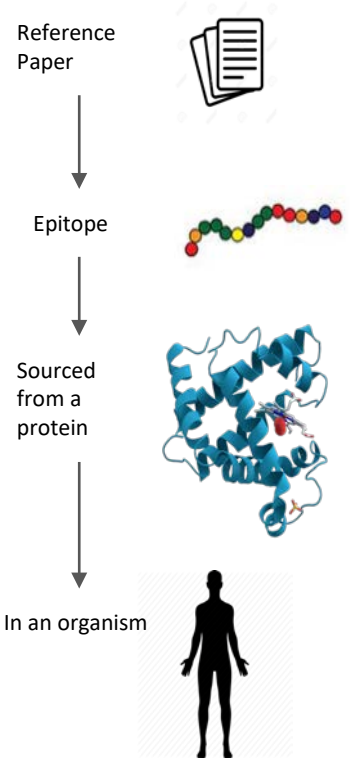


diamond

MMseqs2

Use Cases for Peptide Matching

Curation for IEDB



A large peptidome dataset improves HLA class I epitope prediction across most of the human population

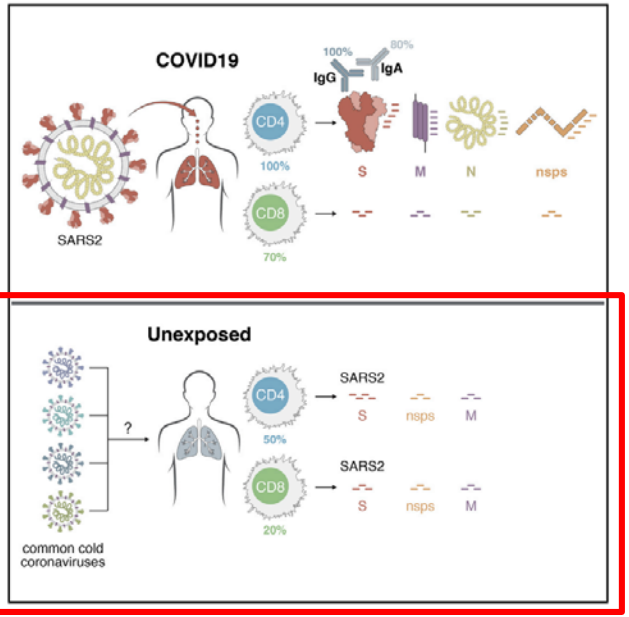
Siranush Sarkizova[1,2,13], Susan Klaeger[2,13], Phuong M. Le[3], Letitia W. Li[3], Giacomo Oliveira[3], Hasmik Keshishian[2], Christina R. Hartigan[2], Wandi Zhang[3], David A. Braun[2,3,4,5], Keith L. Ligon[2,4,6,7], Pavan Bachireddy[2,3,5], Ioannis K. Zervantonakis[8], Jennifer M. Rosenbluth[8], Tamara Ouspenskaia[2], Travis Law[2], Sune Justesen[9], Jonathan Stevens[10], William J. Lane[4,10], Thomas Eisenhaure[2], Guang Lan Zhang[3,4,11], Karl R. Clauser[2], Nir Hacohen[2,3,12*], Steven A. Carr[2*], Catherine J. Wu[2,3,4,5*] and Derin B. Keskin[2,3,4,5,11*]

Prediction of HLA epitopes is important for the development of cancer immunotherapies and vaccines. However, current prediction algorithms have limited predictive power, in part because they were not trained on high-quality epitope datasets covering a broad range of HLA alleles. To enable prediction of endogenous HLA class I-associated peptides across a large fraction of the human population, we used mass spectrometry to profile >185,000 peptides eluted from 95 HLA-A, -B, -C and -G mono-allelic cell lines. We identified canonical peptide motifs per HLA allele, unique and shared binding submotifs across alleles and distinct motifs associated with different peptide lengths. By integrating these data with transcript abundance and peptide processing, we developed HLAthena, providing allele-and-length-specific and pan-allele-pan-length prediction models for endogenous peptide presentation. These models predicted endogenous HLA class I-associated ligands with 1.5-fold improvement in positive predictive value compared with existing tools and correctly identified >75% of HLA-bound peptides that were observed experimentally in 11 patient-derived tumor cell lines.

Use Cases for Peptide Matching

Targets of T Cell Responses to SARS-CoV-2 Coronavirus in Humans with COVID-19 Disease and Unexposed Individuals

Graphical Abstract



Authors

Alba Grifoni, Daniela Weiskopf, Sydney I. Ramirez, ..., Davey M. Smith, Shane Crotty, Alessandro Sette

Correspondence

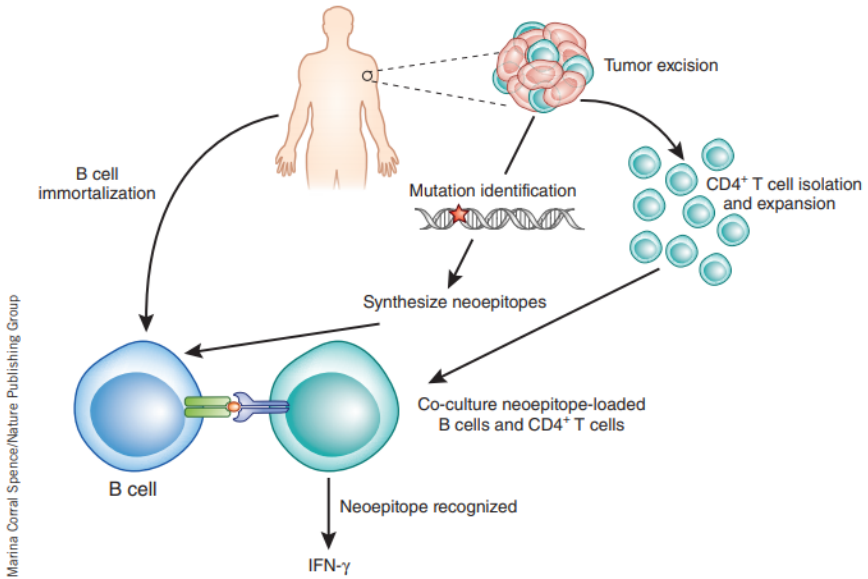
shane@lji.org (S.C.),
alex@lji.org (A.S.)

In Brief

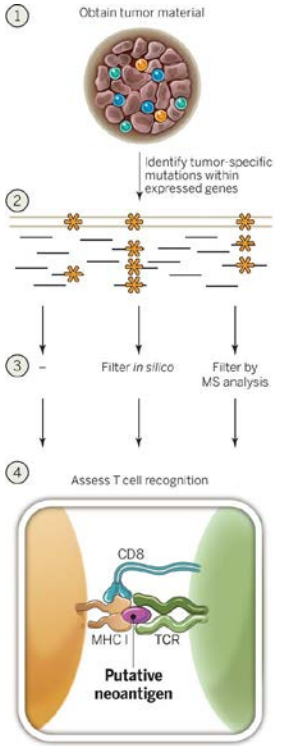
An analysis of immune cell responses to SARS-CoV-2 from recovered patients identifies the regions of the virus that is targeted and also reveals cross-reactivity with other common circulating coronaviruses

Use Cases for Peptide Matching

Neoepitope Similarity For Cancer Vaccine Candidates



Overwijk, W., et al. *Nat Med* (2015)



Schumacher TN, et al. Science. (2015)

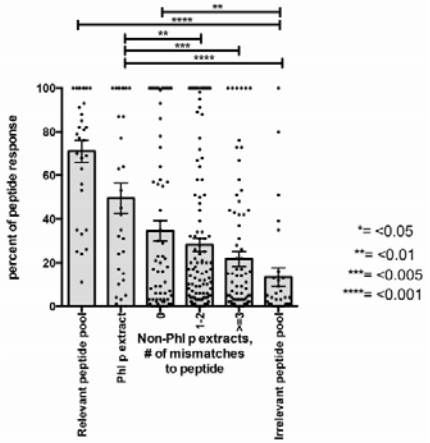
Use Cases for Peptide Matching

Conservation Analyses

- Milk allergens → do conserved peptides in human proteome drive stronger reactions?
- Fungus allergens → does conservation across allergen species determine immunogenicity?
- COVID-19 vaccination → does the spike protein encoded by the mRNA vaccines have similar enough homology that could lead to autoimmune myocarditis?

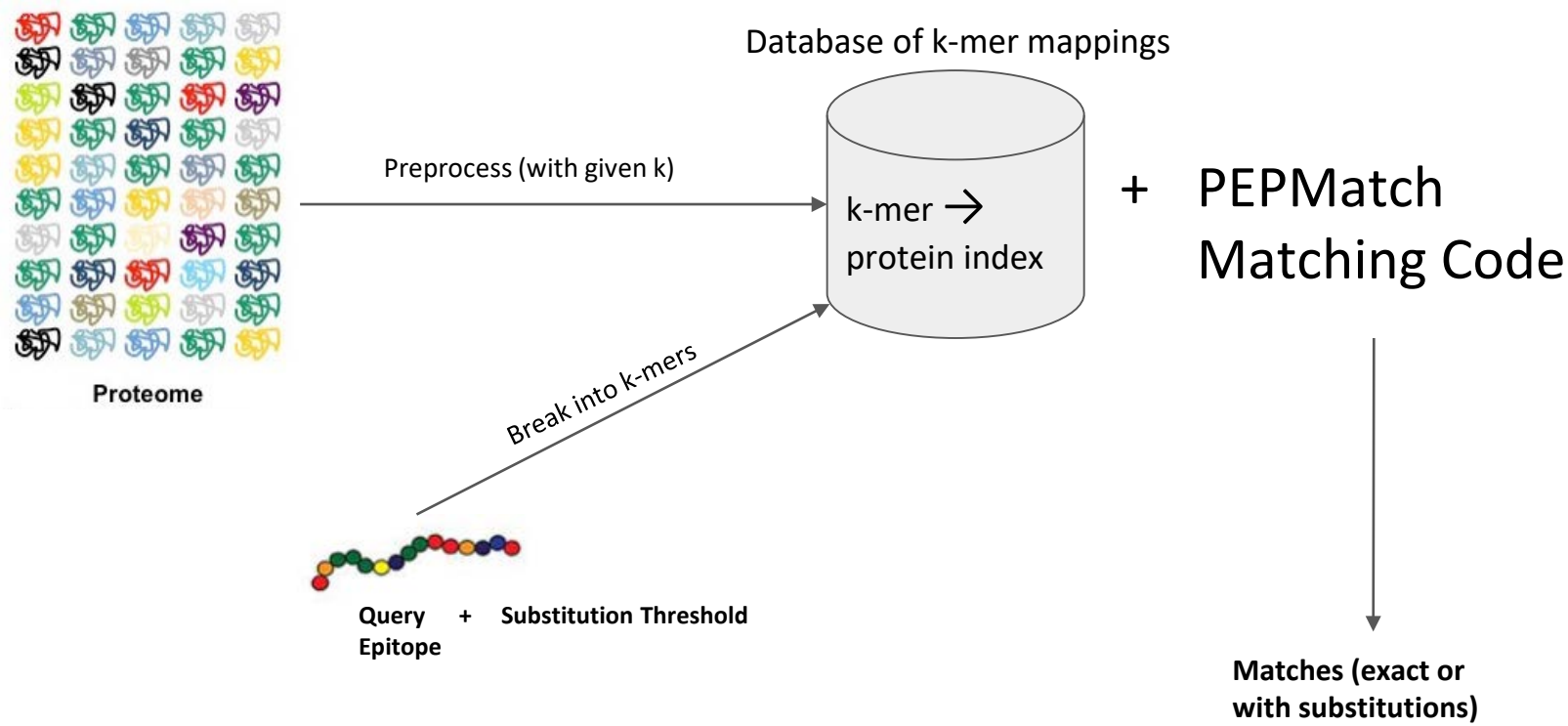
T-cell epitope conservation across allergen species is a major determinant of immunogenicity

Luise Westernberg[1], Véronique Schulten[1], Jason A Greenbaum[1], Sara Natali[3], Victoria Tripple[1], Denise M. McKinney[1], April Frazier[1], Heidi Hofer[2], Michael Wallner[2], Federica Sallusto[3,4], Alessandro Sette[1], and Bjoern Peters[1]

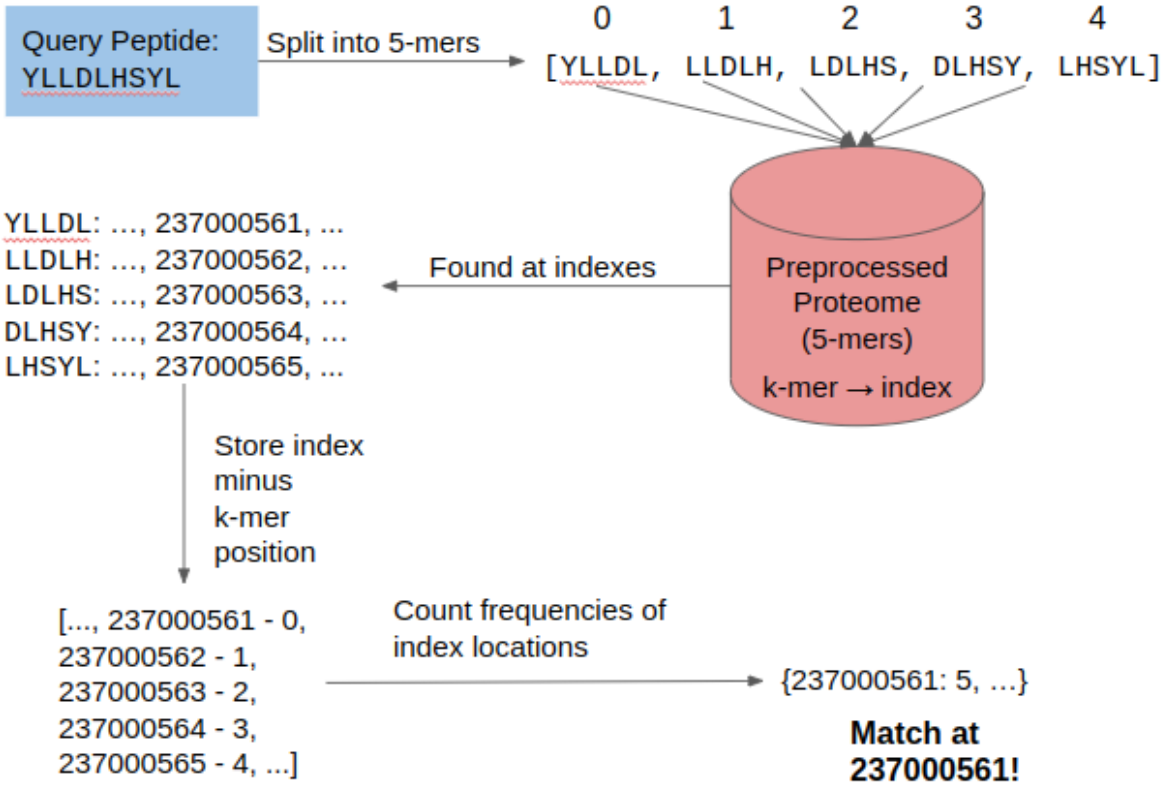


Westernberg, Luise et al. Journal of Allergy and Clinical Immunology (2016)

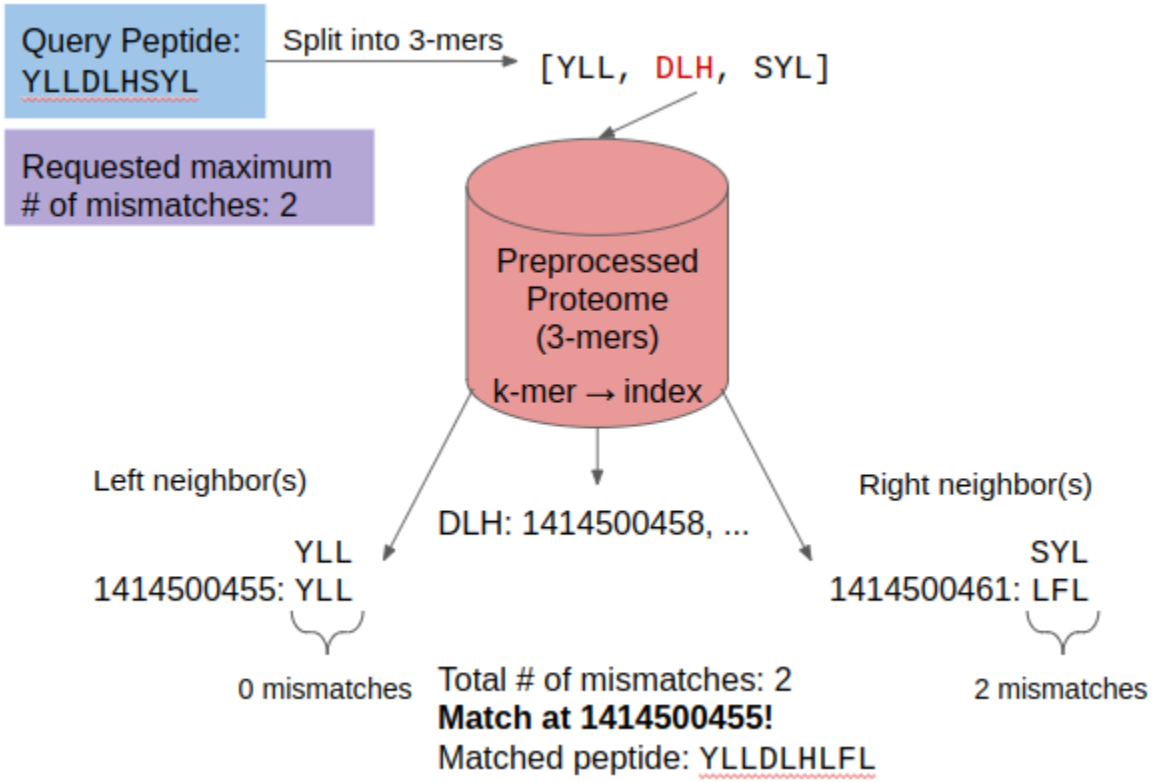
PEPMatch: Tool Overview



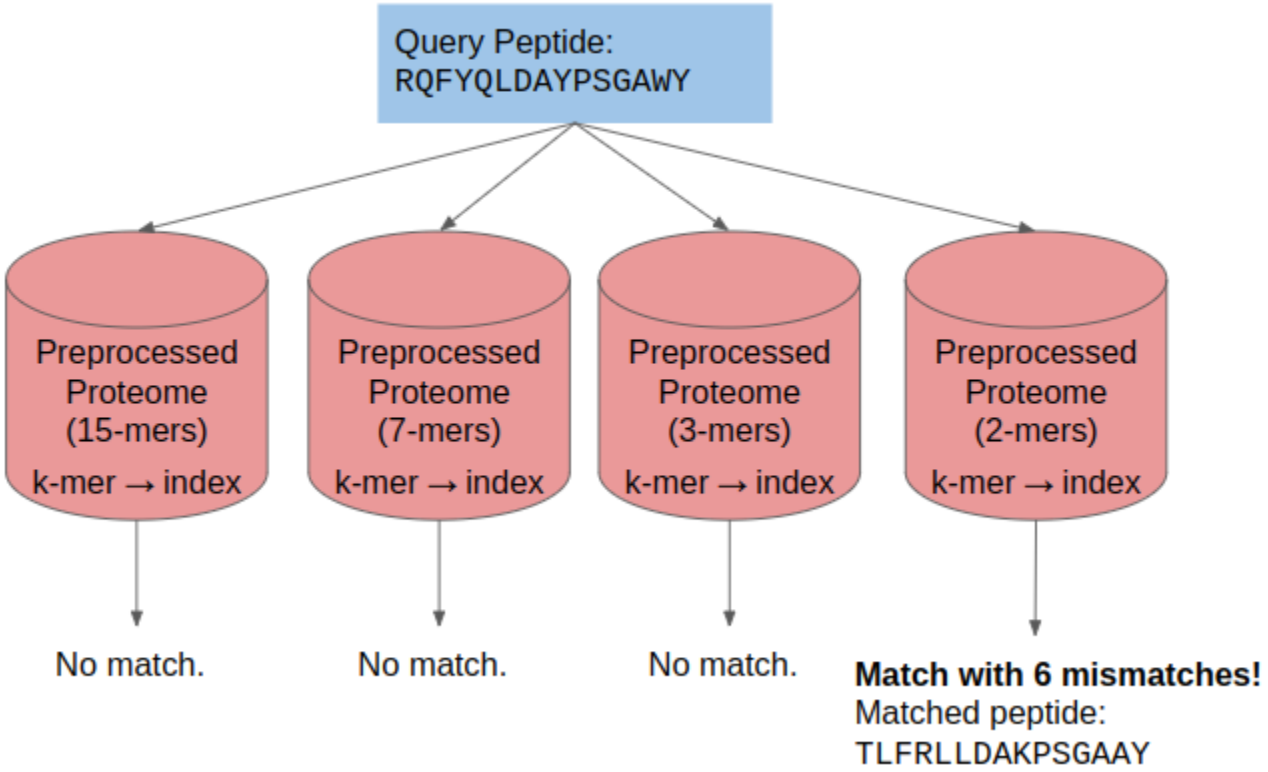
PEPMatch: Exact Match Search



PEPMatch: Searching with Substitutions (Mismatches)



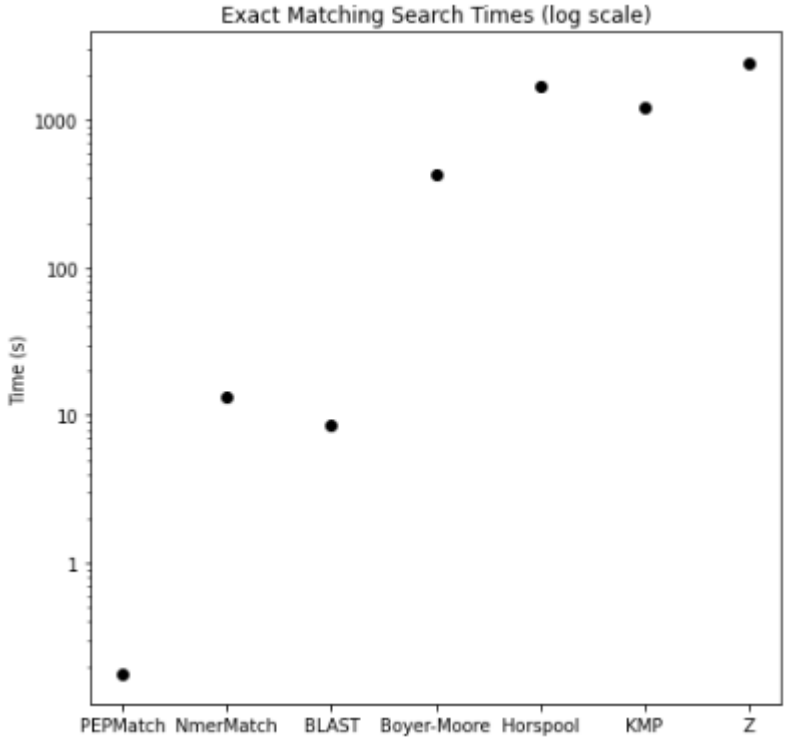
PEPMatch: Best Match Search



Benchmarking: Exact Matching Search

Task: Find 1,000 9-mers in the human proteome.

Algorithm/Tool	Proteome Preprocessing Time (s)	Query Preprocessing Time (s)	Searching Time (s)	Total Time (s)	Accuracy (%)
PEPMatch	36.6	N/A	0.18	36.8	100
NmerMatch	28.7	0.003	13.2	41.9	100
BLAST	0.53	N/A	8.48	9.01	99.4
Horspool	N/A	N/A	423.7	423.7	100
Boyer-Moore	N/A	N/A	1700.6	1700.6	100
Knuth-Morris-Pratt (KMP)	N/A	N/A	1204.2	1204.2	100
Z	N/A	N/A	2448.4	2448.4	100



PEPMatch outperformed BLAST and other string searching algorithms.

* excluded DIAMOND and MMseqs2 due to low accuracy

Benchmarking: Searching with Substitutions (Mismatches)

Task: Find 628 peptides of various lengths (8-15) in all betacoronavirus proteomes with 2 substitutions or less.

Tool	Proteome Preprocessing Time (s)	Query Preprocessing Time (s)	Searching Time (s)	Total Time (s)	Accuracy (%)
PEPMatch	40.7	N/A	51.9	92.6	100
NmerMatch	280.7	0.003	25.2	305.9	100
BLAST	0.61	N/A	156.2	156.8	73.4
DIAMOND	0.25	N/A	6.38	6.63	6.34
MMseqs2	3.76	N/A	1.14	4.90	7.34

PEPMatch and NmerMatch outperformed BLAST in speed and accuracy. They outperformed DIAMOND and MMseqs2 in accuracy by a lot. NmerMatch takes longer to preprocess.

Benchmarking: Best Match Search

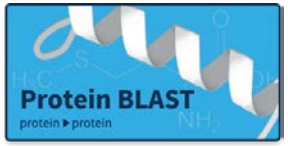
Task: Find the best match of 111 15-mers in the human proteome. Some matches may have 7-8 mismatches.

Tool	Proteome Preprocessing Time (s)	Query Preprocessing Time (s)	Searching Time (s)	Total Time (s)	Accuracy (%)
PEPMatch	45.2	N/A	155.4	200.6	100
NmerMatch	52.02	0.001	356.1	410.1	100
BLAST	0.56	N/A	18.21	18.77	24.3
DIAMOND	0.18	N/A	4.24	4.42	12.4
MMseqs2	2.98	N/A	0.64	3.62	15.0

PEPMatch outperformed NmerMatch in speed and BLAST, DIAMOND and MMseqs2 in accuracy.

Benchmarking Framework

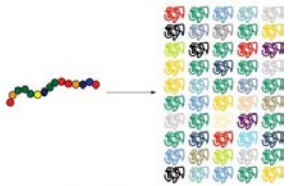
Your Tool
(any language)



diamond

Datasets

MMseqs2



PEPMatch

Expected output, time logging and accuracy.

https://github.com/IEDB/PEPMatch

Using PEPMatch

Command Line

```
[dan@foxp3 ~]$ pip install pepmatch
> pepmatch-preprocess -p 9606.fasta
> -k 5 -f sql
```

```
[dan@foxp3 ~]$ pepmatch-match -q peptides.fasta
> -p 9606 -m 0 -k 5
```

Python Module

```
1  #!/usr/bin/env python3
2
3  from pepmatch import Preprocessor, Matcher
4
5  Preprocessor('human_proteome.fasta', k=5, preprocess_format='sql').preprocess()
6  Matcher('peptides.fasta', 'human_proteome', k=5, max_mismatches=0).match()
```

Soon to be on the IEDB!

Thank you
for listening!