大连理工大学
DALIAN UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF LEICESTER

Peptide Detectability Prediction Based on Interpretable Classification Model

Junjie Dong

Applied Chemistry

DLI, Dalian University of Technology

Supervisor: Zengyou He

May 21, 2022

Outline

❶ Introduction

Background

Research Purpose and Motivation

State-of-the-art

Contribution

❷ Interpretable Peptide Detectability Prediction Model

Workflow of Model

Sequential Patterns Mining Module

Decision Rule Set Learning Module

❸ Experiment and Discussion

Setup

Results and Evaluation

❹ Reference

Reference

Outline

1. **Introduction**
   - Background
   - Research Purpose and Motivation
   - State-of-the-art
   - Contribution
2. **Interpretable Peptide Detectability Prediction Model**
   - Workflow of Model
   - Sequential Patterns Mining Module
   - Decision Rule Set Learning Module
3. **Experiment and Discussion**
   - Setup
   - Results and Evaluation
4. **Reference**
   - Reference

Background

Proteomics[1]

Obtain protein information about cells, tissues and organisms.

Significance:

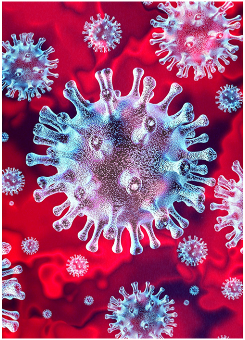


Figure 1:
Disease mechanisms.

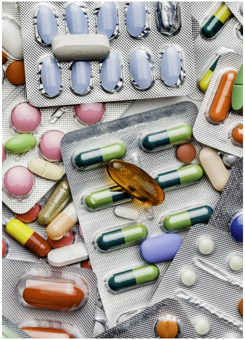


Figure 2:
Drug discovery.

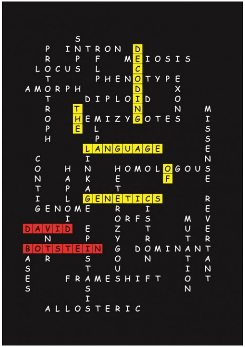


Figure 3:
Genetic language.

Purpose and Motivation

Proteomics

- Protein identification and inference
- **“Bottom-up”** and “Top-down”
- How do we infer protein correctly?
- Necessary to ensure the peptide detectability!

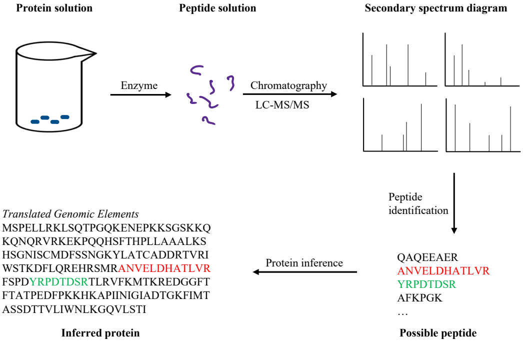
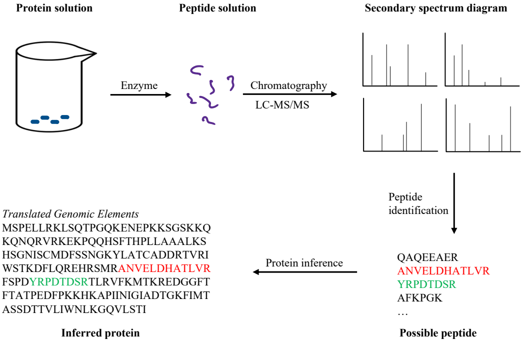

Protein solution
Peptide solution
Secondary spectrum diagram
Enzyme
Chromatography
LC-MS/MS
Peptide identification
Protein inference
Translated Genomic Elements
MSPELLRKLSQTPGQKENEPKKSGSKKQ
KQNQRVRKEKPQQHSFTHPLLAAALKS
HSGNISCMDFSSNGKYLATCADDRTVRI
WSTKDFLQREHRSMRANVELDHATLVR
FSPDYRPDTDSRTLRVFKMTKREDGGFT
FTATPEDFPKKHKAPIINIGIADTGKFIMT
ASSDTTVLIWNLKGQVLSTI
Inferred protein
QAQEEAER
ANVELDHATLVR
YRPDTDSR
AFKPGK
…
Possible peptide

Figure 4: A “bottom-up” approach for protein identification.

Purpose and Motivation

Proteomics

- Protein identification and inference
- **“Bottom-up”** and “Top-down”
- How do we infer protein correctly?
- Necessary to ensure the peptide detectability!



Protein solution

Enzyme

Peptide solution

Chromatography

LC-MS/MS

Secondary spectrum diagram

Peptide identification

QAQEEAER
ANVELDHATLVR
YRPDTDSR
AFKPGK
…

Possible peptide

Protein inference

Translated Genomic Elements
MSPELLRKLSQTPGQKENEPKKSGSKKQ
KQNQRVRKEKPQQHSFTHPLLAAALKS
HSGNISCMDFSSNGKYLATCADDRTVRI
WSTKDFLQREHRSMRANVELDHATLVR
FSPDYRPDTDSRTLRVFKMTKREDGGFT
FTATPEDFPKKHKAPIINIGIADTGKFIMT
ASSDTTVLIWNLKGQVLSTI

Inferred protein

Figure 4: A “bottom-up” approach for protein identification.

State-of-the-art

Mainstream method

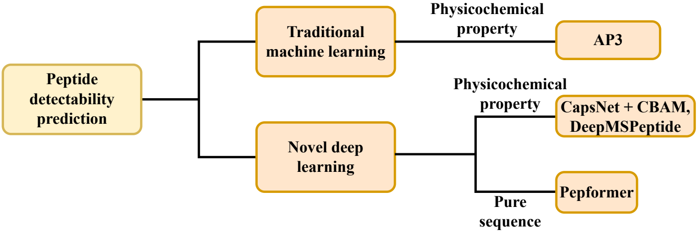


Figure 5: Categories for mainstream method.

Disadvantage on these model:

- Traditional machine learning e.g. AP3[2].
  - Rely on prior knowledge and featurization
- Novel deep learning, e.g. Pepformer[3], DeepMSPeptide[4].
  - High computing resources

Not interpretable!

State-of-the-art

Mainstream method

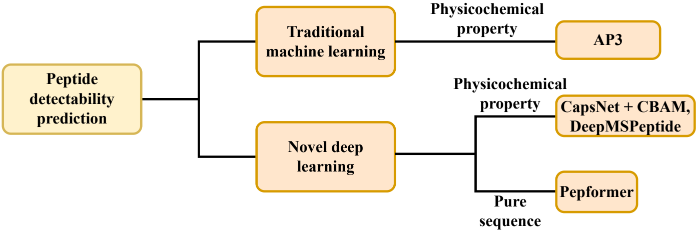


Figure 5: Categories for mainstream method.

Disadvantage on these model:

- Traditional machine learning e.g. AP3[2].
  - Rely on prior knowledge and featurization
- Novel deep learning, e.g. Pepformer[3], DeepMSPeptide[4].
  - High computing resources

Not interpretable!

Contribution

Why interpretbility is necessary[5]?

1. Promote trust in the model
   - Understanding the decision step
   - Debugged and audited
2. Human curiosity and learning
   - Scientific purpose
3. Ascertain the mechanism of peptide detection.
   - Improve the experimental procedure.

This project presented a **interpretable** peptide detectability prediction model.

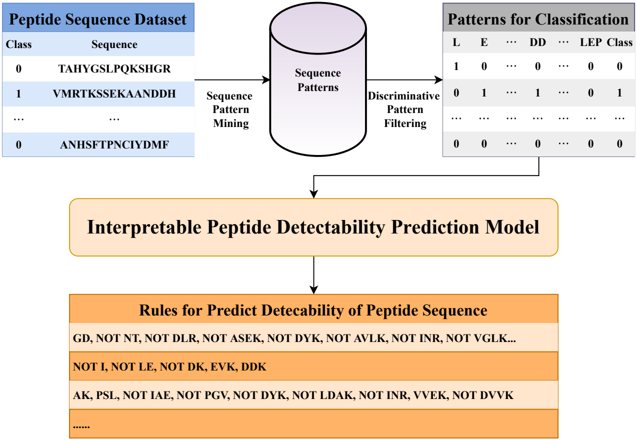
Outline

1. **Introduction**
   - Background
   - Research Purpose and Motivation
   - State-of-the-art
   - Contribution
2. **Interpretable Peptide Detectability Prediction Model**
   - Workflow of Model
   - Sequential Patterns Mining Module
   - Decision Rule Set Learning Module
3. **Experiment and Discussion**
   - Setup
   - Results and Evaluation
4. **Reference**
   - Reference

Model Structure

Main structure

- Sequential Patterns Mining Module
- Decision Rule Set Learning Module



Peptide Sequence Dataset

Class	Sequence
0	TAHYGSLPQKSHGR
1	VMRTKSSEKAANDDH
…	…
0	ANHSFTPNCIYDMF

Patterns for Classification

L	E	⋯	DD	⋯	LEP	Class
1	0	⋯	0	⋯	0	0
0	1	⋯	1	⋯	0	1
⋯	⋯	⋯	⋯	⋯	⋯	⋯
0	0	⋯	0	⋯	0	0

Interpretable Peptide Detectability Prediction Model

Rules for Predict Detecability of Peptide Sequence
GD, NOT NT, NOT DLR, NOT ASEK, NOT DYK, NOT AVLK, NOT INR, NOT VGLK...
NOT I, NOT LE, NOT DK, EVK, DDK
AK, PSL, NOT IAE, NOT PGV, NOT DYK, NOT LDAK, NOT INR, VVEK, NOT DVVK
......

Sequential Patterns

A sequence is an ordered list of symbols.

A peptide sequence is a combination of itemset of amino acid list in the table below.

Example: AERANVELDH

Table 1: Twenty kinds of amino acids.

Amino acid	Abbreviation	Sign	Amino acid	Abbreviation	Sign
Alanine	Ala	A	Leucine	Leu	L
Arginine	Arg	R	Lysine	Lys	K
Asparagine	Asn	N	Methionine	Met	M
Aspartic acid	Asp	D	Phenylalanine	Phe	F
Cysteine	Cys	C	Proline	Pro	P
Glutamic acid	Glu	E	Serine	Ser	S
Glutamine	Gln	Q	Threonine	Thr	T
Glycine	Gly	G	Tryptophan	Trp	W
Histidine	His	H	Tyrosine	Tyr	Y
Isoleucine	Ile	I	Valine	Val	V

Goal: finding all subsequences that appear frequently in a sequence database. Example: $\langle AE\rangle$, $\langle AR\rangle$, $\langle ER\rangle$ etc.

k-mer

Extracting patterns based on substring.
Example: $\langle AER \rangle, \langle ERA \rangle, \langle RAN \rangle$ etc. in AERANVELDH

Advantage:

- Strict location information
- Simply implement.

Algorithm 2.1: k**-mers** $(String\ Seq, Integer\ k)$

```
$L \leftarrow length(seq)$
k$-mers \leftarrow$ new array of $L - k + 1$ empty strings
for $n \leftarrow to\ L - k$ do
    $k - mers[n] \leftarrow$ subsequence of seq from letter $n$ inclusive to letter $n + k$ exclusive
end for
return $k - mers$
```

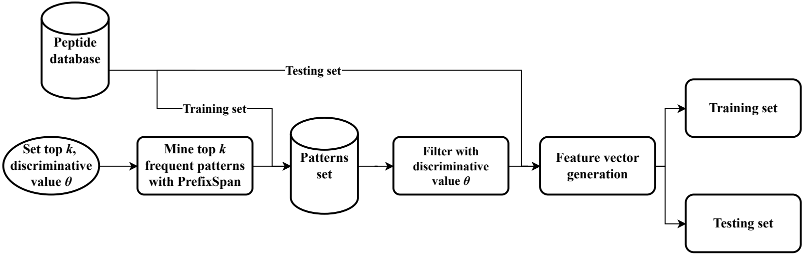
PrefixSpan[6]

Advantage:

- Only consider patterns *exist* in the database.
- uses the concept of **database projection** and **depth-first search**.

Procedure:

1. Select top-k frequent patterns
2. Filter patterns
    - Redundancy
    - Difficult for the training process of next module.


Peptide database
Testing set
Training set
Set top k, discriminative value θ
Mine top k frequent patterns with PrefixSpan
Patterns set
Filter with discriminative value θ
Feature vector generation
Training set
Testing set

Discriminant Patterns Filter

Definition 2.1: Discriminant value[7]

$$Disc(s, D) = \frac{Occ(s, D_{positive})}{|D_{positive}|} - \frac{Occ(s, D_{negative})}{|D_{negative}|}, \quad (1)$$

where $Disc(s, D)$ refers to the discriminant ability of pattern s to positive and negative classes in a database D.

Algorithm 2.2: contrast$(P_{pos}$,P_{neg},**threshold** θ, $D)$

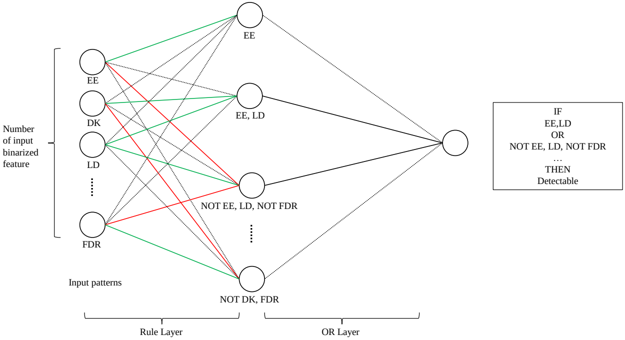
```
res ← new list
for all $i \in P_{pos}$ and $P_{pos}$ do
    if $Disc(i, D) >= \theta$ then
        res.append($i$)
    end if
end for
return res
```

Decision Rule Network[8]

Main workflow in decision rule network

Input: feature vector generated from sequential patterns mining module

Output: rules set for predicting peptide detectability



Number of input binarized feature

EE

DK

LD

FDR

Input patterns

EE

EE, LD

NOT EE, LD, NOT FDR

NOT DK, FDR

IF
EE,LD
OR
NOT EE, LD, NOT FDR
…
THEN
Detectable

Rule Layer

OR Layer

Figure 6: Workflow of generating rules for peptide prediction.

Outline

1 Introduction

Background

Research Purpose and Motivation

State-of-the-art

Contribution

2 Interpretable Peptide Detectability Prediction Model

Workflow of Model

Sequential Patterns Mining Module

Decision Rule Set Learning Module

3 Experiment and Discussion

Setup

Results and Evaluation

4 Reference

Reference

Parameter Setting-k-mer[9]

k is set to be 3 according to“elbow” method.

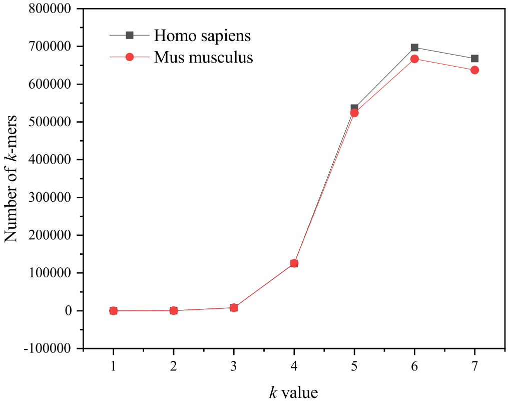
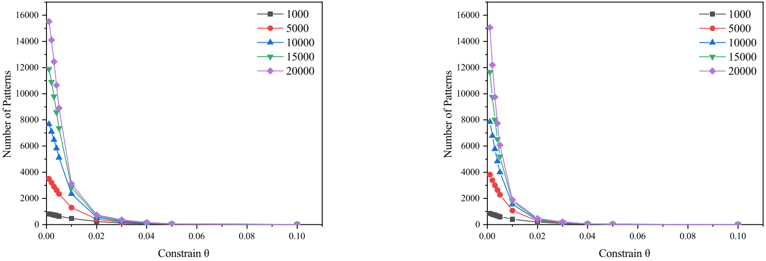


Figure 7: Number of k-mers versus k value.

Parameter Setting: PrefixSpan

Setting top-20000 under threshold 0.02.



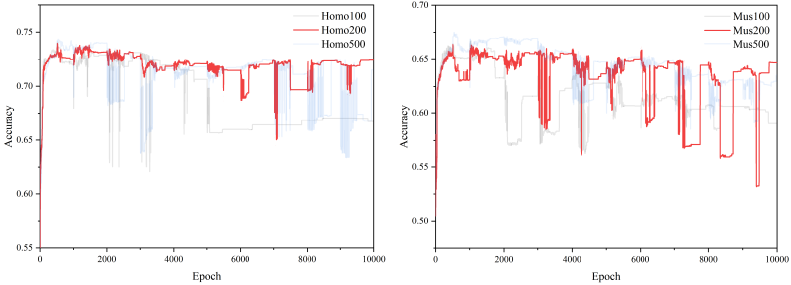
(a) Homo sapiens,

(b) Mus musculus,

Figure 8: Number of patterns versus constrain θ

Parameter Setting: rule number

Number of rule number is set to be 200.



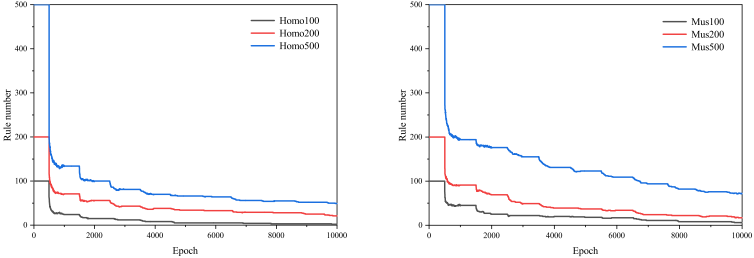
(a) Homo sapiens,

(b) Mus musculus,

Figure 9: Accuracy versus epoch

Parameter Determination: rule number

Number of rule number is set to be 200.



(a) Homo sapiens,

(b) Mus musculus,

Figure 10: Rule number in decision set versus epoch.

Results and Evaluation

Table 2: Experiment results on Homo sapiens.

Models	ACC	SP	SN	MCC
iBCM+RF	0.5767	0.6190	0.5215	0.1573
AP3 [1]	0.6416	0.5949	0.6881	0.2843
2-mer-DRN	0.6800	**0.7664**	0.5973	0.3692
SeqDT	0.7151	0.7089	0.7213	0.4345
PrefixSpan-DRN	0.7201	**0.7873**	0.6470	0.4414
PepFormer [1]	0.8066	0.7213	0.8915	0.6221

[1] The comparison result are from the work of Pepformer[3].

Table 3: Experiment results on Mus musculus.

Models	ACC	SP	SN	MCC
iBCM+RF	0.5767	0.6190	0.5215	0.1573
2-mer-DRN	0.5956	**0.8864**	0.3047	0.2349
PrefixSpan-DRN	0.6447	**0.8127**	0.4799	0.3099
AP3[1]	0.6462	0.5993	0.6928	0.2934
SeqDT	0.6537	0.6467	0.6568	0.3035
PepFormer[1]	0.7521	0.6421	0.8629	0.5176

[1] The comparison result are from the work of Pepformer[3].

Details of Rules in Decision Set

Table 4: Details of rules in decision sets.

Models	Rule numbers	Rule length	N. conditions[1]	P. conditions[2]	Accuracy
Homo100	6	10.33	50	12	0.6483
Homo200	17	11.76	168	32	0.7201
Homo500	49	10.53	372	144	0.7150
Mus100	2	5.5	7	4	0.5184
Mus200	22	8.45	130	56	0.6447
Mus500	94	14.11	453	309	0.6050

[1] Negative conditions.

[2] Positive conditions.

Outline

1 **Introduction**
- Background
- Research Purpose and Motivation
- State-of-the-art
- Contribution

2 **Interpretable Peptide Detectability Prediction Model**
- Workflow of Model
- Sequential Patterns Mining Module
- Decision Rule Set Learning Module

3 **Experiment and Discussion**
- Setup
- Results and Evaluation

4 **Reference**
- Reference

Reference I

[1] WASINGER V C, CORDWELL S J, CERPA-POLJAK A, et al. Progress with Gene-Product Mapping of the Mollicutes: Mycoplasma Genitalium[J]. Electrophoresis, 1995, 16(1): 1090-1094.

[2] GAO Z, CHANG C, YANG J, et al. AP3: An Advanced Proteotypic Peptide Predictor for Targeted Proteomics by Incorporating Peptide Digestibility[J]. Analytical Chemistry, 2019, 91(13): 8705-8711.

[3] CHENG H, RAO B, LIU L, et al. PepFormer: End-to-End Transformer-Based Siamese Network to Predict and Enhance Peptide Detectability Based on Sequence Only[J]. Analytical Chemistry, 2021, 93(16): 6481-6490.

[4] SERRANO G, GURUCEAGA E, SEGURA V. DeepMSPeptide: Peptide Detectability Prediction Using Deep Learning[J]. Bioinformatics, 2020, 36(4): 1279-1280.

[5] MOLNAR C. Interpretable Machine Learning[M]. 2021.

[6] Jian Pei, Jiawei Han, MORTAZAVI-ASL B, et al. PrefixSpan,: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth[C]. in: Proceedings 17th International Conference on Data Engineering. Heidelberg, Germany: IEEE Comput. Soc, 2001: 215-224.

Reference II

[7] HE Z, ZHANG S, WU J. Significance-Based Discriminative Sequential Pattern Mining[J]. Expert Systems with Applications, 2019, 122: 54-64.

[8] QIAO L. Learning Accurate and Interpretable Decision Rule Sets from Neural Networks[C]. in: 35th AAAI Conference on Artificial Intelligence. 2020.

[9] DEOROWICZ S, GUDYŚ A, DŁUGOSZ M, et al. Kmer-Db: Instant Evolutionary Distance Estimation[J]. Bioinformatics (Oxford, England), 2019, 35(1): 133-136.

Thanks!