Relative Distance Protein Fingerprint (RD-PFP) Algorithm Combined with Machine Learning for Searching DNA Mimic Proteins

相對距離蛋白質指紋 (RD-PFP) 演算法結合機器學習尋找 DNA 擬態蛋白質

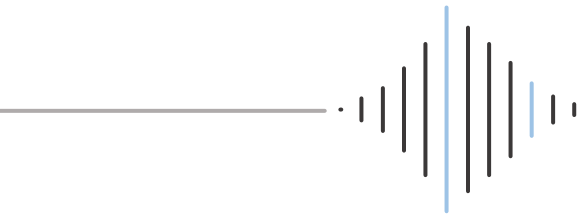
Student: Chia-Yen Chien (簡嘉妍)
Advisor: Prof. Sun-Yuan Hsieh (謝孫源 教授)

國立成功大學資訊工程研究所
Department of Computer Science and Information Engineering,
National Cheng Kung University

July 10, 2023

OUTLINE

- Introduction
- Methods
- Experiments and Results
- Discussion
- Conclusion



Introduction

- DNA-binding Protein
- DNA Double Helix Structure
- DNA Mimic Protein
- Literature Review and Limitation
- Proposed Method

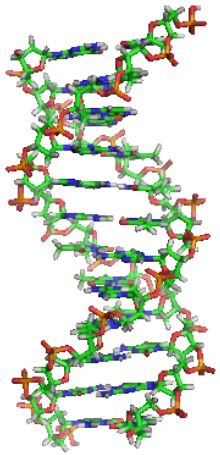
INTRODUCTION

DNA-binding Protein

International Networks and High-performance Computation Lab

- Proteins with a **DNA-binding domain**.

[1]



- Include:
  - Transcription factors that regulate the transcription process
  - Various polymerases
  - Nucleases that cleave DNA molecules
  - Histones involved in the packaging and transcription of chromosomes in the nucleus

- **DNA-binding domain** can "read" a specific DNA sequence, which is the **outer surface** of the DNA double helix.

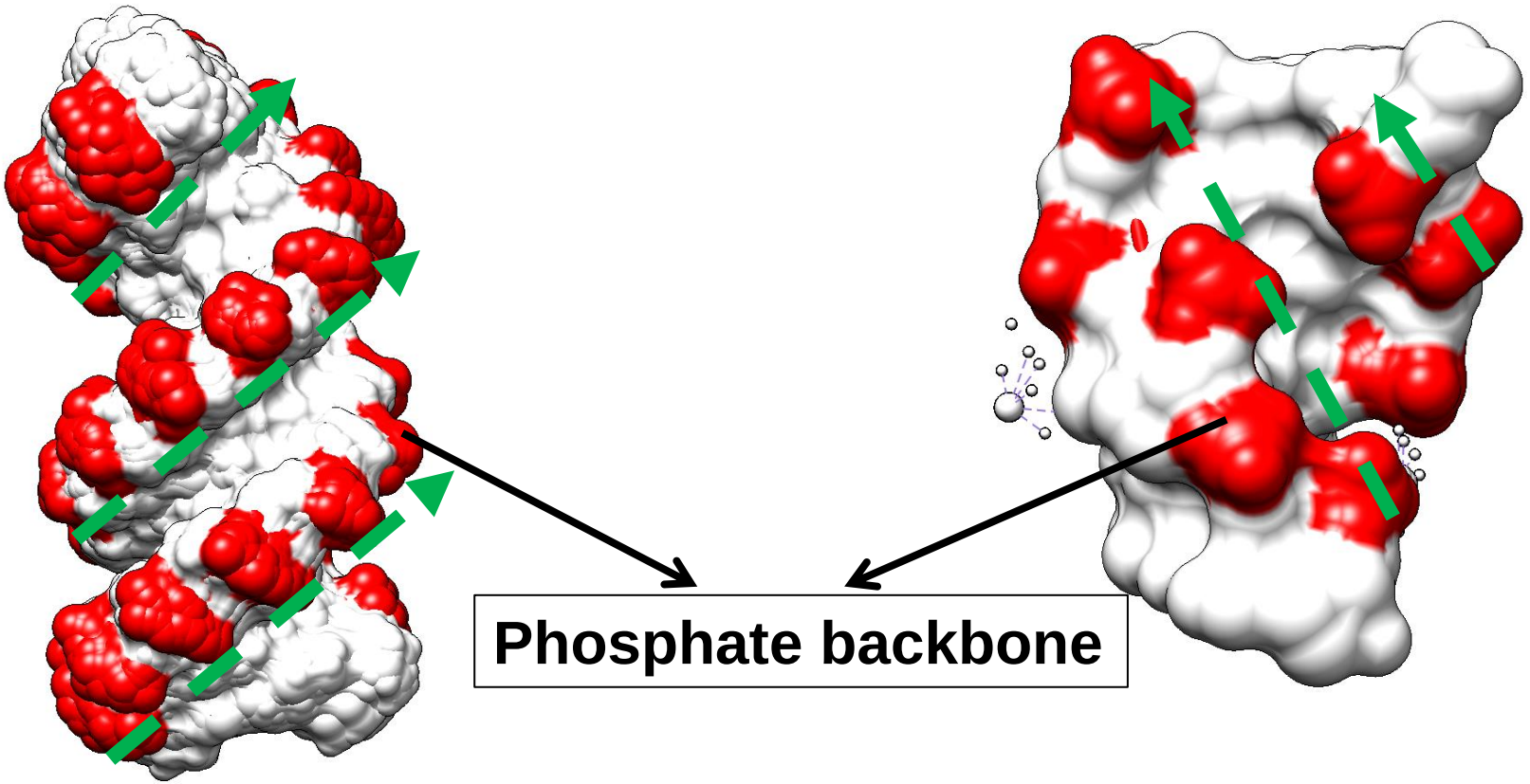
[1] https://zh.wikipedia.org/zh-tw/File:DNA_orbit_animated.gif

INTRODUCTION

DNA Double Helix Structure

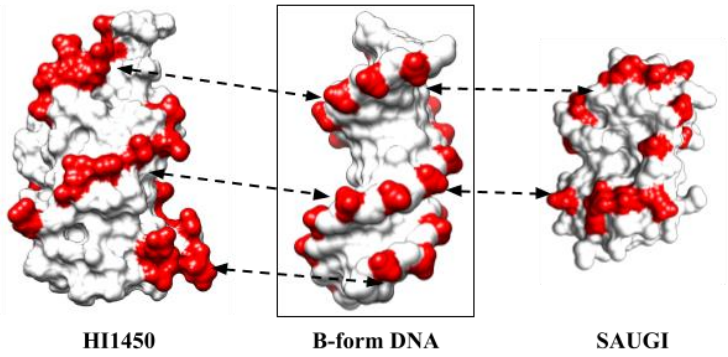
- **B-form DNA**
  Right hand spin

- **Z-form DNA**
  Left hand spin



INTRODUCTION

DNA Mimic Protein



- Mimic DNA phosphate backbone by using **negatively charged amino acids**, namely aspartic acid (ASP or D) and glutamic acid (GLU or E).

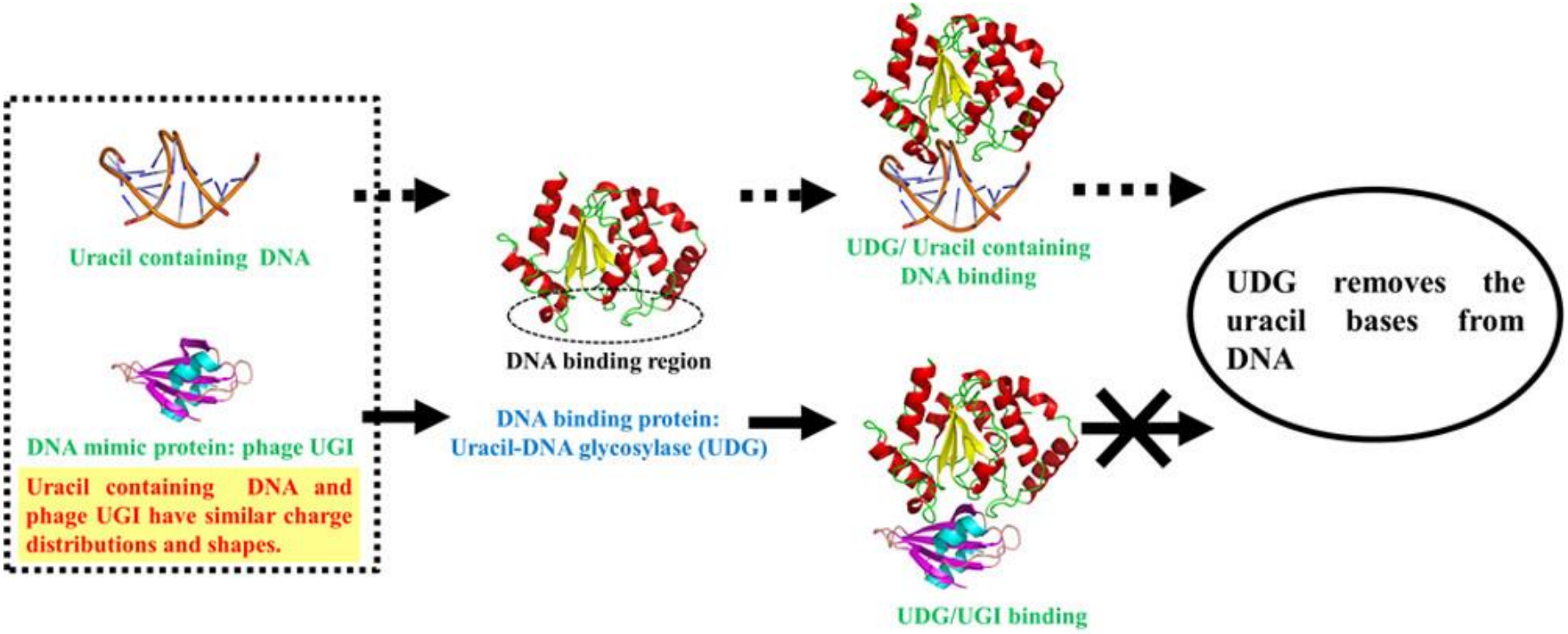
- **Directly occupying the DNA-binding domain** on the DNA-binding protein.

- DNA mimic proteins are typically **D/E-rich**. Specifically, 15%~20% of amino acids are negatively charged.

INTRODUCTION

DNA Mimic Protein

- An example of how the phage **UGI DNA mimic** protein can act as a competitive inhibitor for UDG.



① [1] **DNA Mimic Proteins: Functions, Structures, and Bioinformatic Analysis** Hao-Ching Wang, Chun-Han Ho, Kai-Cheng Hsu, Jinn-Moon Yang, and Andrew H.-J. Wang *Biochemistry* **2014** *53* (18), 2865-2874DOI: 10.1021/bi5002689

INTRODUCTION

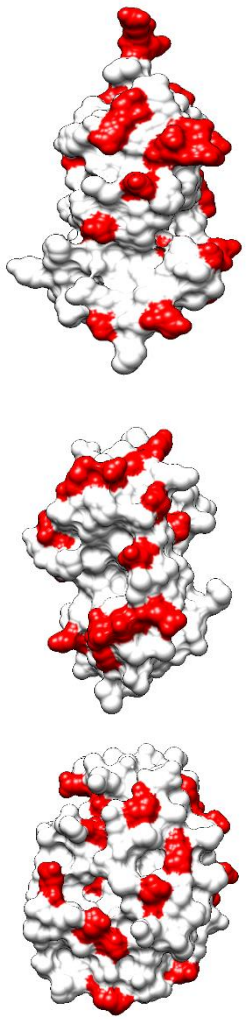
Literature Review and Limitation

International

ce

☐ The application of DNA mimic proteins in biotechnology

Author / Years	DNA mimic protein	Function
[1] MD Walkinshaw / 2002 [2] Artem Isaev / 2020 [3] Guoqiang Zhang / 2012	**Ocr**	Restriction
[4] Hao-Ching Wang / 2013 [5] Hao-Ching Wang / 2016 [6] Yi-Ting Liao / 2020 [7] Joshua P Ramsay / 2016	**SAUGI**	DNA repair Transfer
[8] Hao-Ching Wang / 2012 [9] Ming-Fen Huang / 2017	**DMP19**	Transcription



[1] MD Walkinshaw, P Taylor, SS Sturrock, C Atanasiu, T Berge, Robert M Henderson, JM Edwardson, and DTF Dryden. Structure of ocr from bacteriophage t7, a protein that mimics b-form dna. Molec
[2] Artem Isaev, Alena Drobiazko, Nicolas Sierro, Julia Gordeeva, Ido Yosef, Udi Qimron, Nikolai V Ivanov, and Konstantin Severinov. Phage t7 dna mimic protein ocr is a potent inhibitor of brex defenc
[3] Guoqiang Zhang, Wenzhao Wang, Aihua Deng, Zhaopeng Sun, Yun Zhang, Yong Liang, Yongsheng Che, and Tingyi Wen. A mimicking -of-dnamethylation-patterns pipeline for overcoming the resti
[4] Hao-Ching Wang, Kai-Cheng Hsu, Jinn-Moon Yang, Mao-Lun Wu, TzuPing Ko, Shen-Rong Lin, and Andrew H-J Wang. Staphylococcus aureus protein saugi acts as a uracil-dna glycosylase inhibito
[5] Hao-Ching Wang, Chun-Han Ho, Chia-Cheng Chou, Tzu-Ping Ko, MingFen Huang, Kai-Cheng Hsu, and Andrew H-J Wang. Using structural-based 62 protein engineering to modulate the differentia
[6] Yi-Ting Liao, Shin-Jen Lin, Tzu-Ping Ko, Chang-Yi Liu, Kai-Cheng Hsu, and Hao-Ching Wang. Structural insight into the differential interactions between the dna mimic protein saugi and two gamma
[7] Joshua P Ramsay. Replicating methicillin resistance? Nature Structural & Molecular Biology, 23(10):874–875, 2016
[8] Hao-Ching Wang, Tzu-Ping Ko, Mao-Lun Wu, Shan-Chi Ku, Hsing-Ju Wu, and Andrew H-J Wang. Neisseria conserved protein dmp19 is a dna mimic protein that prevents dna binding to a hypotheti
[9] Ming-Fen Huang, Shin-Jen Lin, Tzu-Ping Ko, Yi-Ting Liao, Kai-Cheng Hsu, and Hao-Ching Wang. The monomeric form of neisseria dna mimic protein dmp19 prevents dna from binding to the histon

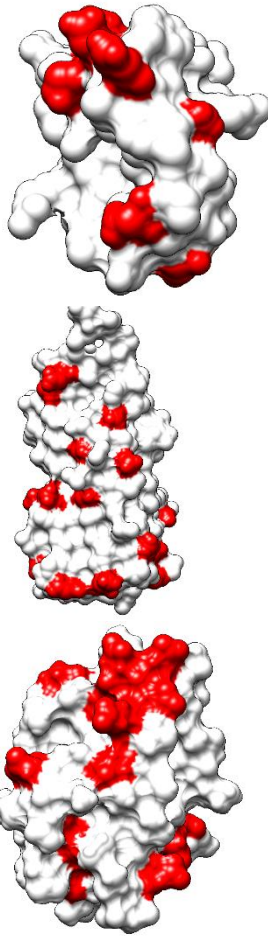
INTRODUCTION

Literature Review and Limitation

International Networks and High-performance Computation Lab

☐ **The application of DNA mimic proteins in biotechnology**

Author / Years	DNA mimic protein	Function
[1] Ying Wu / 2016	**VPR**	Transfer
[2] Subray S Hegde / 2005	**Mfpa**	Topology
[3] Jiyung Shin / 2017	**AcrIIA4**	Restriction

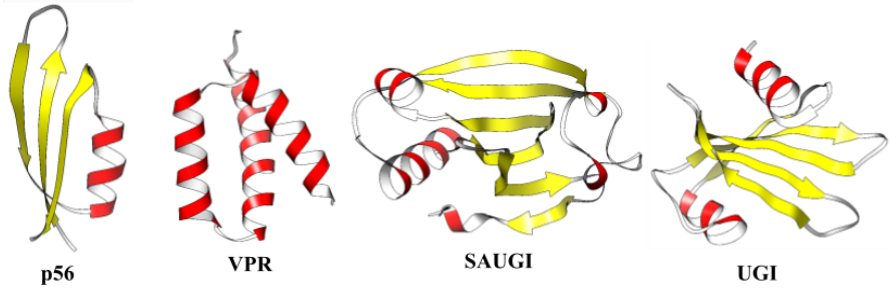


[1] Ying Wu, Xiaohong Zhou, Christopher O Barnes, Maria DeLucia, Aina E Cohen, Angela M Gronenborn, Jinwoo Ahn, and Guillermo Calero. The ddb1–dcaf1–vpr–ung2 crystal structure reveals how hiv -1 vpr steers human ung2 toward destruction. Nature structural & molecular biology, 23(10):933–940, 2016.
[2] Subray S Hegde, Matthew W Vetting, Steven L Roderick, Lesley A Mitchenall, Anthony Maxwell, Howard E Takiff, and John S Blanchard. A 63 fluoroquinolone resistance protein from mycobacterium tuberculosis that mimics dna. Science, 308(5727):1480 –1483, 2005.
[3] Jiyung Shin, Fuguo Jiang, Jun-Jie Liu, Nicolas L Bray, Benjamin J Rauch, Seung Hyun Baik, Eva Nogales, Joseph Bondy-Denomy, Jacob E Corn, and Jennifer A Doudna. Disabling cas9 by an anti-crispr dna mimic. Science advances, 3(7):e1701620, 2017.

INTRODUCTION

Literature Review and Limitation

- The amino acid sequence and structural appearance of DNA mimic proteins are not similar **only a few DNA mimic proteins have been published.**
- The DNA mimic proteins p56, Vpr, SAUGI and UGI inhibit UDG activity, but their **structures and sequences are diverse**.
- **Simplifying the process of comparing protein structures** can facilitate comparisons of 3D structures and DNA phosphate backbons.



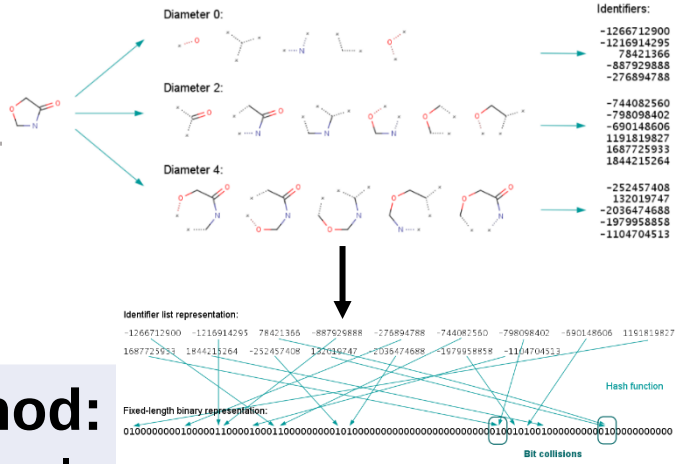

```
P56     -------DSYD--VTMLLQDDDGKQYYEYH-KGLS----LSDFEVLYGNTADEIIKLRLD 46
VPR     ----------------------EWTLELLEELKSEAVRHFPRIWLHNLGQH----I----- 30
SAUGI   MTLELQLKHYITNLFNLPKDEKWECESIEEIADD---ILPDQYV-RLGALSN-KIL---Q 52
UGI     -----------TNLSDIIEKETGKQLVIQE-SIL---MLPEEVEEVIGNKPESDIL---V 42
                                .                      *      *

P56     KVNDFVDSYDVTMLLQDDDGKQYYEYHKGLSLSDFEVLYGNTADEIIKLRLDKVL----- 101
VPR     -YETYGDTWAGVE-----------A-IIRILQQLLFIHFRIGCSH--------------- 62
SAUGI   TYTYYSDTLHE---------SNIYPFILYYQKQLIAIGYIDENHDMDFLYLHNTIMPLLD 103
UGI     -HTAYDESTDENVMLLTSDAPEYKPWALVIQD-------SNGENKIKML---------- 83
            : ::                                .

P56     ------      101
VPR     ------      62
SAUGI   QRYLLT      109
UGI     ------      83
```

INTRODUCTION

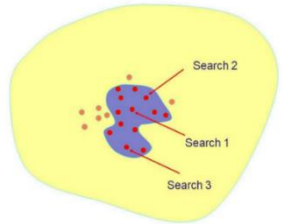
Literature Review and Limitation



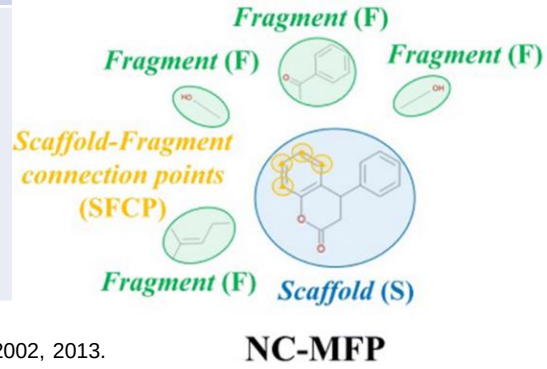
☐ Molecular Fingerprint Methods

[1] Extended-connectivity fingerprint method: circular topological fingerprints designed for molecular characterization, similarity searching, and structure-activity modeling.

[2] Fusing similarity rankings in ligand-based virtual screening: combined a series of different types of similarity measures into molecular fingerprints through data fusion.



[3] Natural compound molecular fingerprinting: converts molecular structures into multibit strings.



[1] David Rogers and Mathew Hahn. Extended -connectivity fingerprints. Journal of chemical information and modeling, 50(5):742 –754, 2010

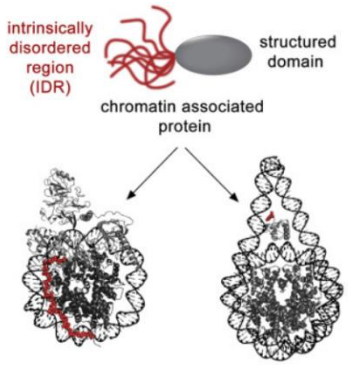
[2] Peter Willett. Fusing similarity rankings in ligand -based virtual screening. Computational and structural biotechnolog y journal, 5(6):e201302002, 2013.

[3] Myungwon Seo, Hyun Kil Shin, Yoochan Myung, Sungbo Hwang, and Kyoung Tai No.
Development of natural compound molecular fingerprint (nc-mfp) with the dictionary of natural products (dnp) for natural productbased drug development. Journal of Cheminformatics, 12(1):1 –17, 202

INTRODUCTION

Literature Review and Limitation

[4]



International Networks and High-performance Computation Lab

☐ Application of Machine Learning to Proteins

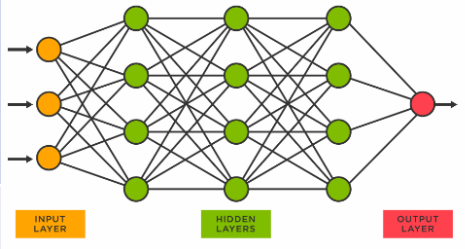
[1] DISpro:

Predict disordered regions based on secondary structure, relative solvent accessibility and one-dimensional recurrent neural networks.

[2] Contact prediction using mutual information and neural nets:

Applied to pairs of **residue positions** and outputs a **probability of contact** between the positions.

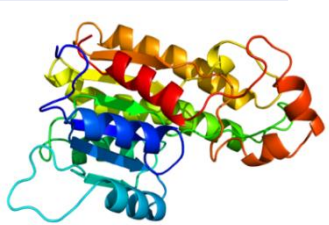
[5]



[3] PROSPECT II:

A threading algorithm which tries to combine all this information in an optimal way for **protein 3D-structure prediction**.

[6]



[1] Jianlin Cheng, Michael J Sweredoski, and Pierre Baldi. Accurate prediction of protein disordered regions by mining protein structure data. Data mining and knowledge discovery, 11:213 –222, 2005.
[2] George Shackelford and Kevin Karplus. Contact prediction using mutual information and neural nets. Proteins: Structure, Function, and Bioinformatics, 69(S8):159–164, 2007.
[3] MyDongsup Kim, Dong Xu, Jun-tao Guo, Kyle Ellrott, and Ying Xu. Prospect ii: protein structure prediction program for genome -scale applications. Protein engineering, 16(9):641–650, 2003.
[4] https://www.sciencedirect.com/science/article/pii/S2589004221000389
[5] https://www.tibco.com/zh-hant/reference-center/what-is-a-neural-network
[6] https://en.wikipedia.org/wiki/Homology_modeling

INTRODUCTION

Proposed Method

- We proposed a new protein fingerprint called the "**Relative Distance Protein Fingerprint** (RD-PFP)" and we employed a **machine learning** method for classification, optimization, and analysis of RD-PFPs.

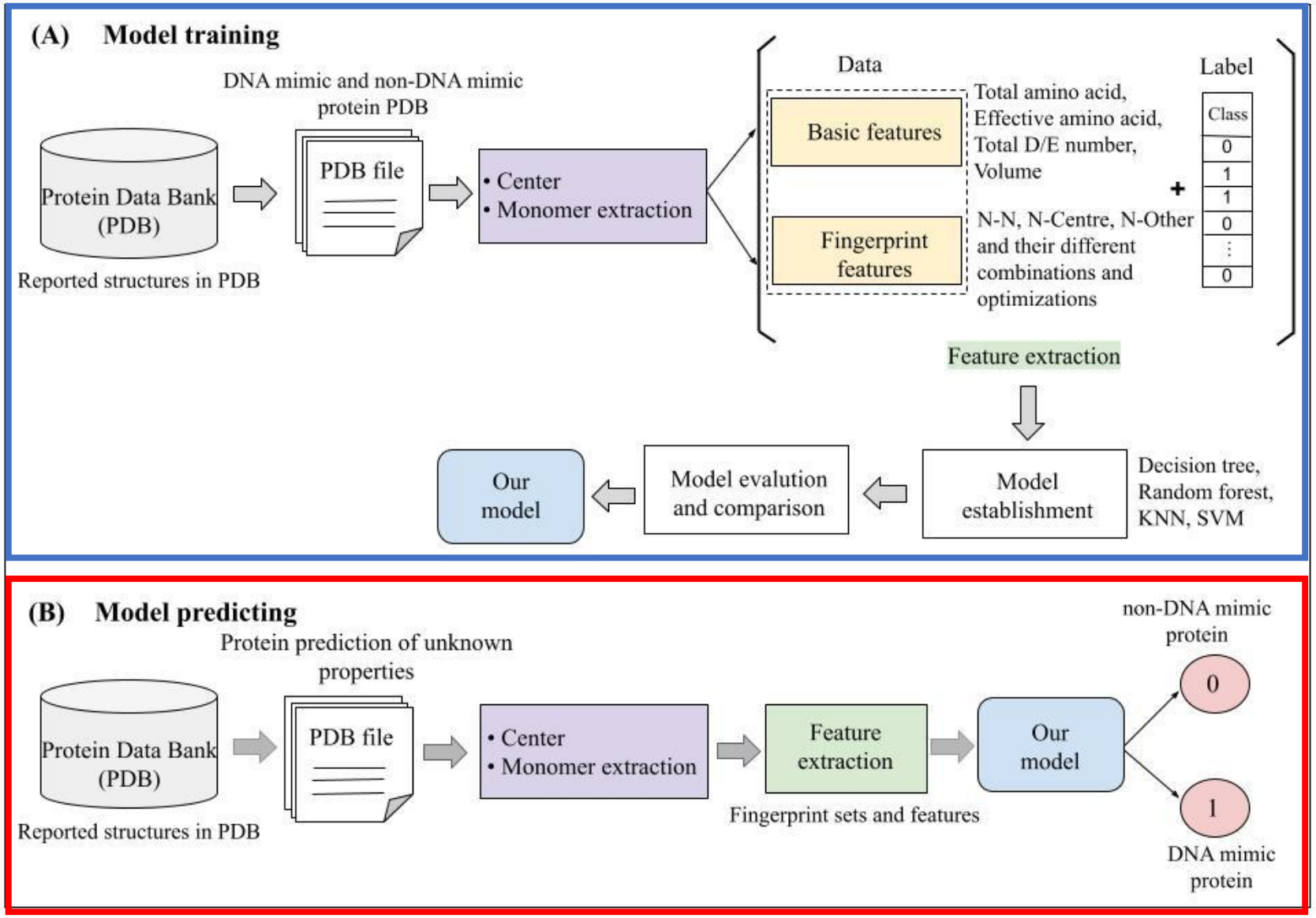
- By utilizing our trained model to analyze proteins with unknown properties, we **identified some potential DNA and not potential DNA mimic proteins**.

Methods

- Flow of Method
- Datasets Preparation
- Extraction of Basic Features
- RD-PFP Algorithm
- Model Candidates
- Metrics

Methods

Flow of Method



Methods

Datasets Preparation

International Networks and High-performance Computation Lab

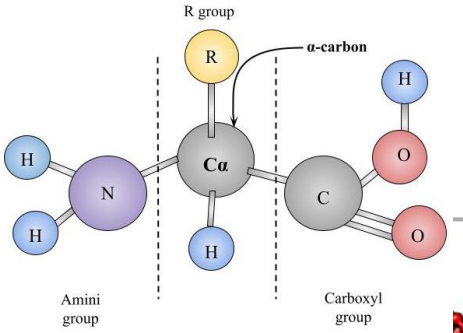
- **DNA mimic proteins**:
  - **Surface negative charge distributions** similar to the DNA phosphate backbone

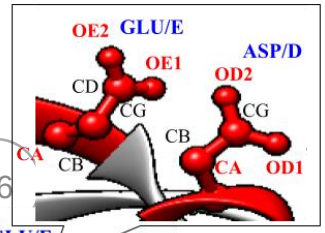
- **Non-DNA mimic proteins:**
  - Sequence identity less than **30%**
  - Amino acid lengths ranging between **100 to 200**
  - Good resolution **(≤2Å)**

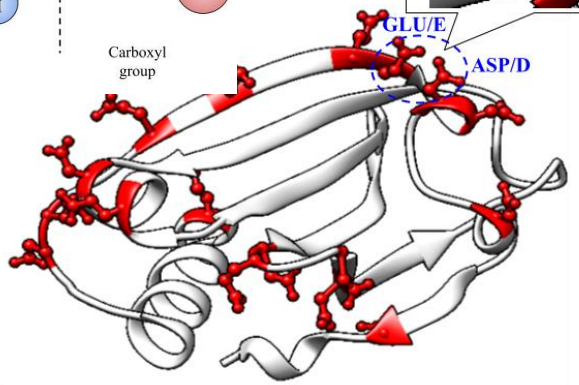
Data Collection	#data
DNA mimic	19
Non-DNA mimic	33
Predict dataset	55

Methods

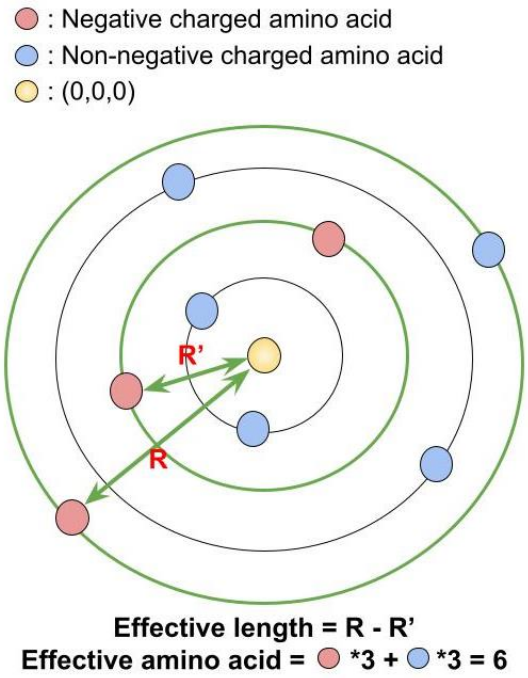
Etraction of Basic Features







- **Total amino acids number**
  - Non-negative: **CA**
  - Negative:
    - **ASP**:**OD1 / OD2** or **GLU**:**OE1 / OE2**
- **Negatively charged amino acids number**
  - ASP / GLU
- **Protein volume**
  - **Effective length^3**
- **Effective amino acid number**



Methods

Etraction of Basic Features

- Basic features extraction of **DNA mimic proteins**.

DNA mimic protein	PDB entry	Total amino acid number	Effective amino acid number	Total D/E number	Volume ($Å^3$)
DMP19	3VJZ	164	155	27	6321.04
HI1450	1NNV	107	102	31	10976.85
GAM	2UUZ	89	82	20	30049.53
Arn	3WX4	98	60	17	8016.96
Ocr	1S7Z	106	92	29	20897.41
MfpA	2BM4	180	148	25	3948.12
NuiA	2O3B	143	122	22	9226.19
AcrF10	6B48	96	92	23	6989.05
DinI	1GHH	81	57	13	1579.50
P56	2LE2	48	44	11	2937.26
AcrF2	5UZ9	92	92	20	17043.81
AcrIIA4	5XBL	87	56	18	1926.53
UGI	1UGI	83	47	18	465.86
ICP11	2ZUG	80	46	15	272.67
Abba	2LZF	68	51	12	1731.46
SAUGI	3WDG	109	85	17	1461.34
CarS	2KSS	86	72	16	5425.69
ArdA	2W82	163	138	37	26619.10
DMP12	3W1O	115	70	19	984.01

Methods

RD-PFP Algorithm

International Networks and High-performance Computation Lab

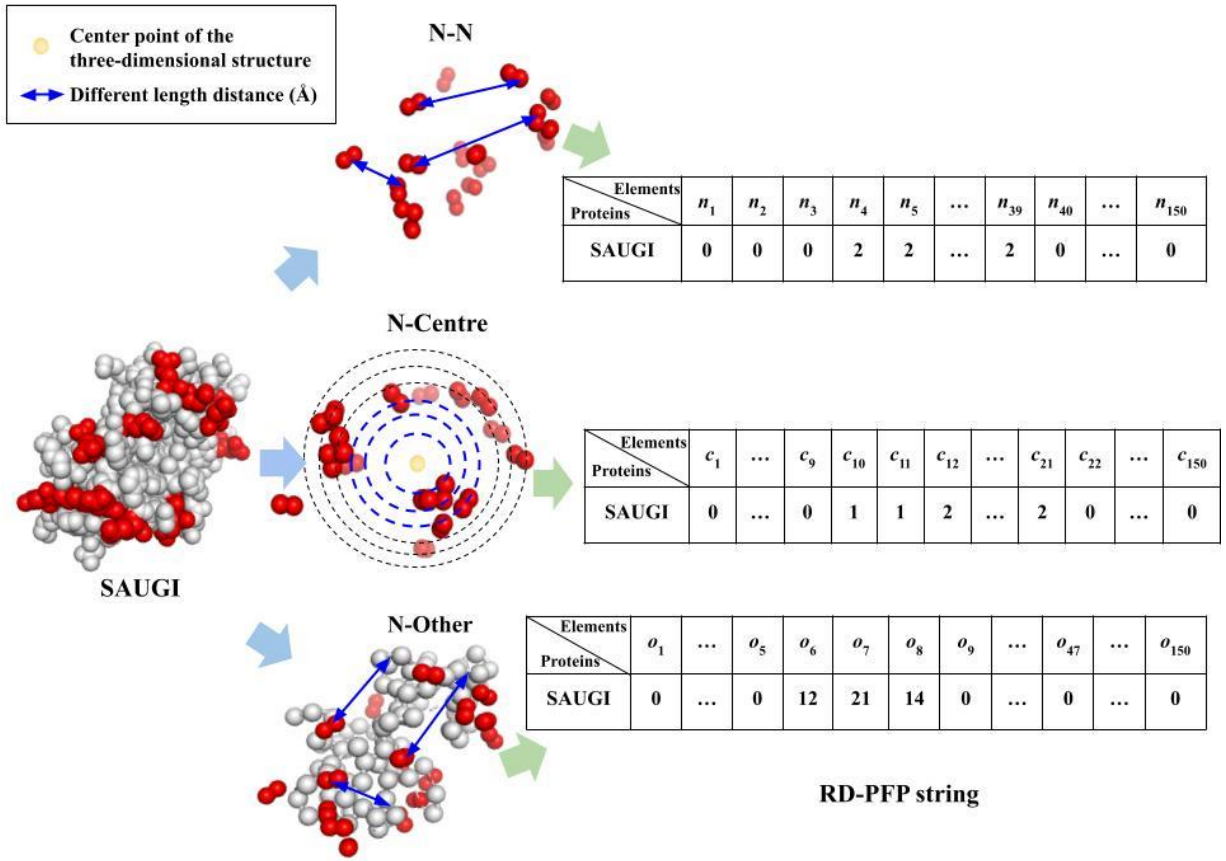
- Three types of protein fingerprint features: **N-N, N-Centre, N-Other.**
- Each set of fingerprints is generated from a **150-dimensional vector**.
- **N-N : (*n*1, *n*2, *n*3, *n*4, ..., *n*150)**
  - ni : number of negatively charged amino acids distance >= i but < (i+1)
- **N-Centre : (*c*1, *c*2, *c*3, *c*4, ..., *c*150)**
  - ci : number of negatively charged amino acids distance from the centroid is >= i but < (i+1)
- **N-Other : (*o*1, *o*2, *o*3, *o*4, ..., *o*150)**
  - oi : number of non-negatively charged amino acids distance from negatively charged amino acids is >= i but < (i+1)
- **Smallest unit of "i" is Ångström (Å).**

Methods

RD-PFP Algorithm

International Networks and High-performance Computation Lab

- **RD-PFP concept**
  - SAUGI is shown.
  - Negatively charged amino acids marked in red.



Proteins \ Elements	n_1	n_2	n_3	n_4	n_5	…	n_{39}	n_{40}	…	n_{150}
SAUGI	0	0	0	2	2	…	2	0	…	0

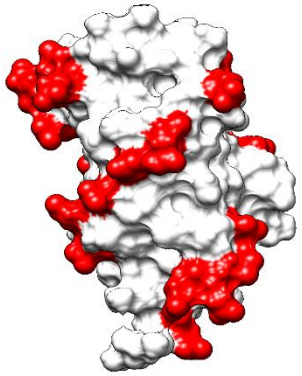
Proteins \ Elements	c_1	…	c_9	c_{10}	c_{11}	c_{12}	…	c_{21}	c_{22}	…	c_{150}
SAUGI	0	…	0	1	1	2	…	2	0	…	0

Proteins \ Elements	o_1	…	o_5	o_6	o_7	o_8	o_9	…	o_{47}	…	o_{150}
SAUGI	0	…	0	12	21	14	0	…	0	…	0

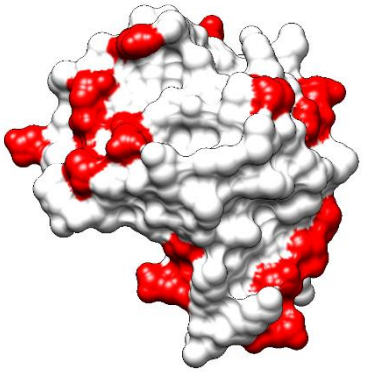
Methods

RD-PFP Algorithm

- **RD-PFP concept**



DNA mimic



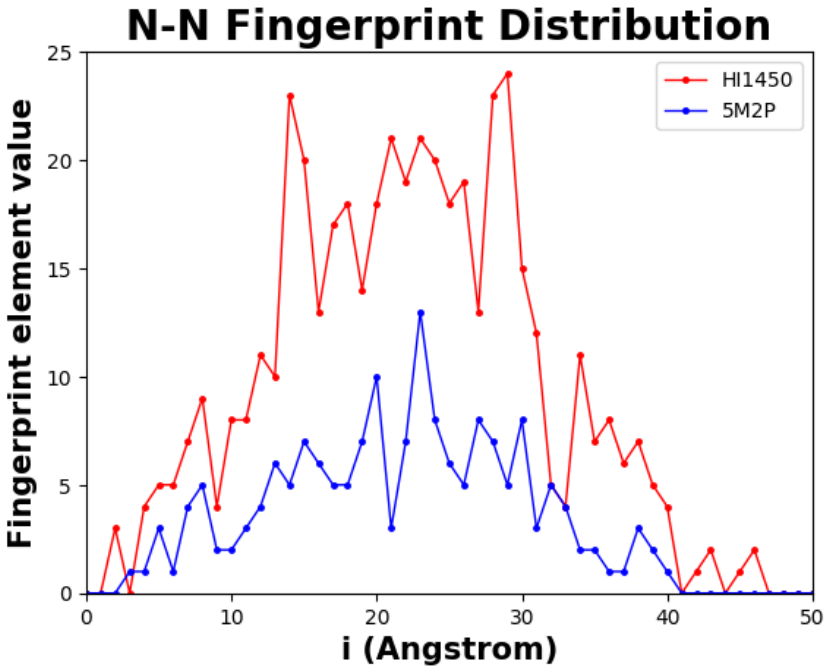
Non-DNA mimic

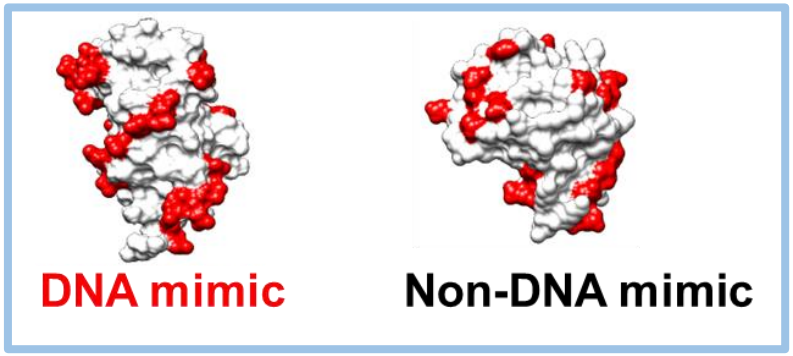
Proteins	HI1450					5M2P				
Total amino acid	107					106				
Elements / Proteins	n_5	n_6	n_7	…	n_{13}	n_{14}	n_{15}	n_{16}	…	n_{47}
HI1450	4	5	5	…	11	10	23	20	…	2
5M2P	1	3	1	…	4	6	5	7	…	0
Elements / Proteins	c_3	c_4	c_5	…	c_{15}	c_{16}	c_{17}	c_{18}	…	c_{26}
HI1450	0	1	0	…	4	3	3	6	…	1
5M2P	0	0	0	…	5	1	2	2	…	0
Elements / Proteins	o_6	o_7	o_8	…	o_{18}	o_{19}	o_{20}	o_{21}	…	o_{48}
HI1450	20	24	29	…	130	130	152	158	…	3
5M2P	15	13	19	…	87	86	119	100	…	0

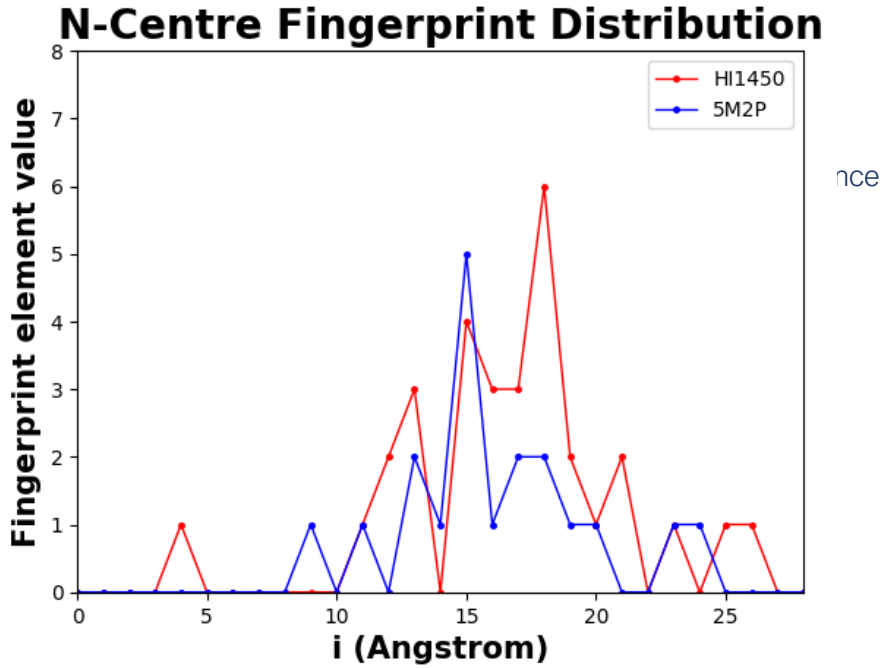
Methods

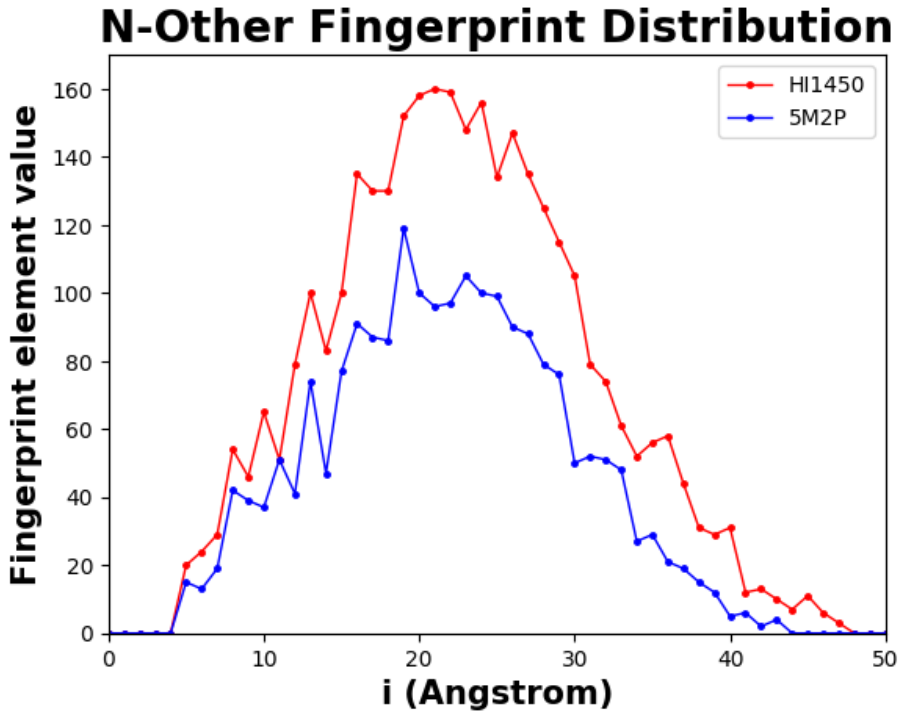
RD-PFP Algorithm

- **RD-PFP concept**







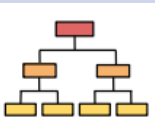
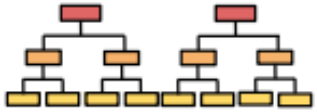
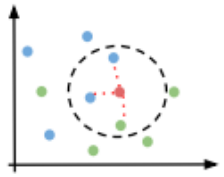
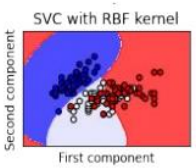


INTRODUCTION

Model Candidates

International Networks and High-performance Computation Lab

☐ Supervised Learning

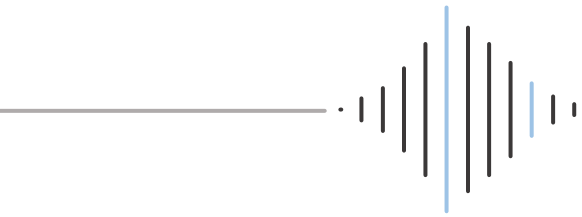
Classifiers	Advantage
[1] [2]Decision trees (DT): Each node represents a decision based on the input variable, each branch represents the outcome of that decision.	• Data preparation easy. • Process data with non-single attributes. 
[3] Random forests (RF): A combination of DT. 	• Mitigates overfitting. • Increases the diversity of the trees.
[4][5] K-nearest neighbors (KNN): Classifies by observing the majority class of its k nearest neighbors in the feature space.	• Unlimited data type. 
[6] [7]Support Vector Machine (SVM): Aims to find the best hyperplane that separates the input data into different classes. 	• Effective to solve the classification problems of high-dimensional features. • Flexibly to solve various nonlinear classification problems.

[1] J. Ross Quinlan. Induction of decision trees. Machine learning, 1:81–106, 1986.
[2] https://zh.wikipedia.org/zh-tw/%E5%86%B3%E7%AD%96%E6%A0%91
[3] Leo Breiman. Random forests. Machine learning, 45:5–32, 2001.
[4] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. Knn model-based approach in classification. In On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings, pages 986–996. Springer, 2003
[5] https://iiac.nccu.edu.tw/at-column/26
[6] Joachims, Thorsten. *Making large-scale SVM learning practical*. No. 1998, 28. Technical report, 1998.
[7] https://shaohua-mi.gitbooks.io/myself/content/svm/svmde-you-dian-que-dian.html

Methods

Metrics

- $\boldsymbol{Precision} = \frac{TP}{TP+FP}$
- $\boldsymbol{Recall} = \frac{TP}{TP+FN}$
- $\boldsymbol{F1\ score} = 2 * \frac{Precision*Recall}{Precision+Recall} = \frac{2*TP}{2*TP+FP+FN}$
- $\boldsymbol{MCC} = \frac{TP*FN-FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}}$
- $\boldsymbol{ACC} = \frac{TP+TN}{TP+TN+FP+FN}$



Experiments and Results

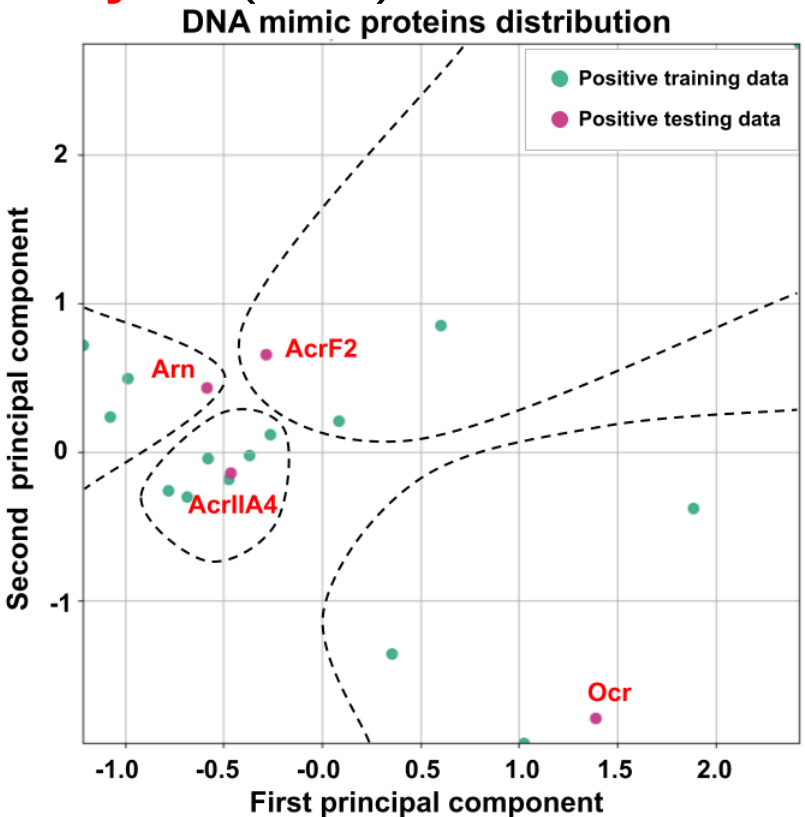
- Data Distribution
- Fingerprint Generation
- Ideal RD-PFP and Basic Features Performance
- Protein Prediction

Experiments and Results

Data Distribution

International Networks and High-performance Computation Lab

- Training and testing sets **8:2**
- **Positive testing set**
  - **N-N fingerprints** obtained from the RD-PFP algorithms and applied **principal components analysis** (PCA)
  - **AcrF2, AcrIIA4, Arn, and Ocr**
- **Negative testing set**
  - Random selection



Experiments and Results

Fingerprint Generation

International Networks and High-performance Computation Lab

- **Baseline RD-PFP (N-N)**
  - **Surface negative charge distribution** in DNA mimic proteins.
  - Necessary to further explore and consider other factors.

Metrics / Classifier	*Precision*	*Recall*	*F_1 score*	*ACC (%)*
DT	0.43	0.49	0.46	56%
RF	0.59	0.55	0.57	60%
KNN	0.51	0.53	0.52	52%
SVM	0.54	0.52	0.53	54%

Experiments and Results

Fingerprint Generation

International Networks and High-performance Computation Lab

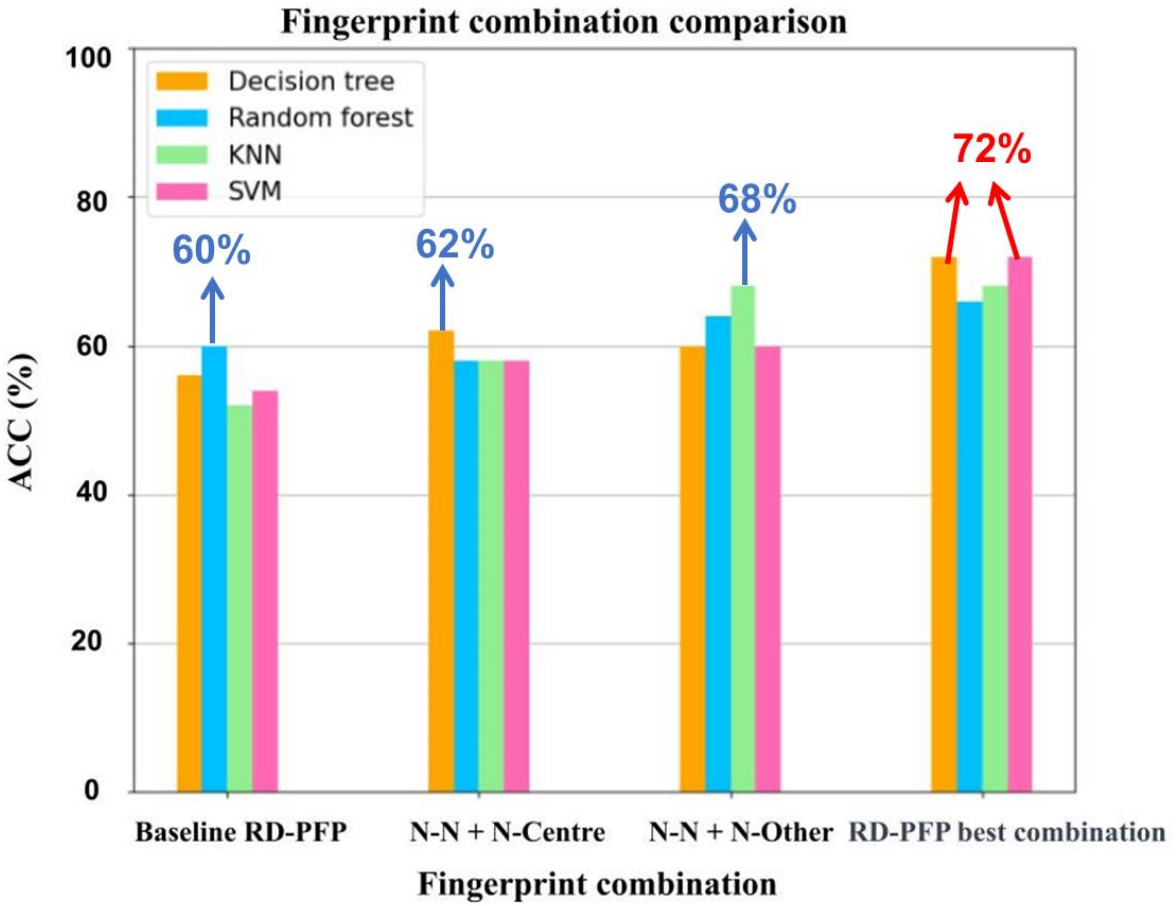
- **RD-PFP Combination**
  - **Comprehensive analysis of the surface negative charge distribution** are more effective in identifying DNA mimic proteins.

Fingerprint combination	N-N + N-Centre			N-N + N-Other			N-N + N-Centre + N-Other		
Metrics / Classifier	Precision	Recall	F_1 *score*	Precision	Recall	F_1 *score*	Precision	Recall	F_1 *score*
DT	0.52	0.52	0.53	0.60	0.58	0.59	**0.66**	**0.56**	**0.61**
RF	0.59	0.57	0.58	0.66	0.63	0.64	**0.69**	**0.62**	**0.65**
KNN	0.54	0.50	0.52	0.51	0.55	0.53	**0.59**	**0.56**	**0.57**
SVM	0.56	0.54	0.55	0.52	0.54	0.53	**0.59**	**0.58**	**0.59**

Experiments and Results

Fingerprint Generation

- **RD-PFP best combination**
  - **N-N + N-Centre + N-Other**



Experiments and Results

Fingerprint Generation

International Networks and High-performance Computation Lab

- **RD-PFP Optimization**
  - DNA mimic proteins tend to have a **high concentration of negatively charged amino acids (D/E rich)** on their surface.
  - **Effective amino acid number** can more accurately represent the number of amino acids than total amino acid number.
  - **Volume** provides a better representation of the **D/E-rich** concept than effective amino acid number parameter.

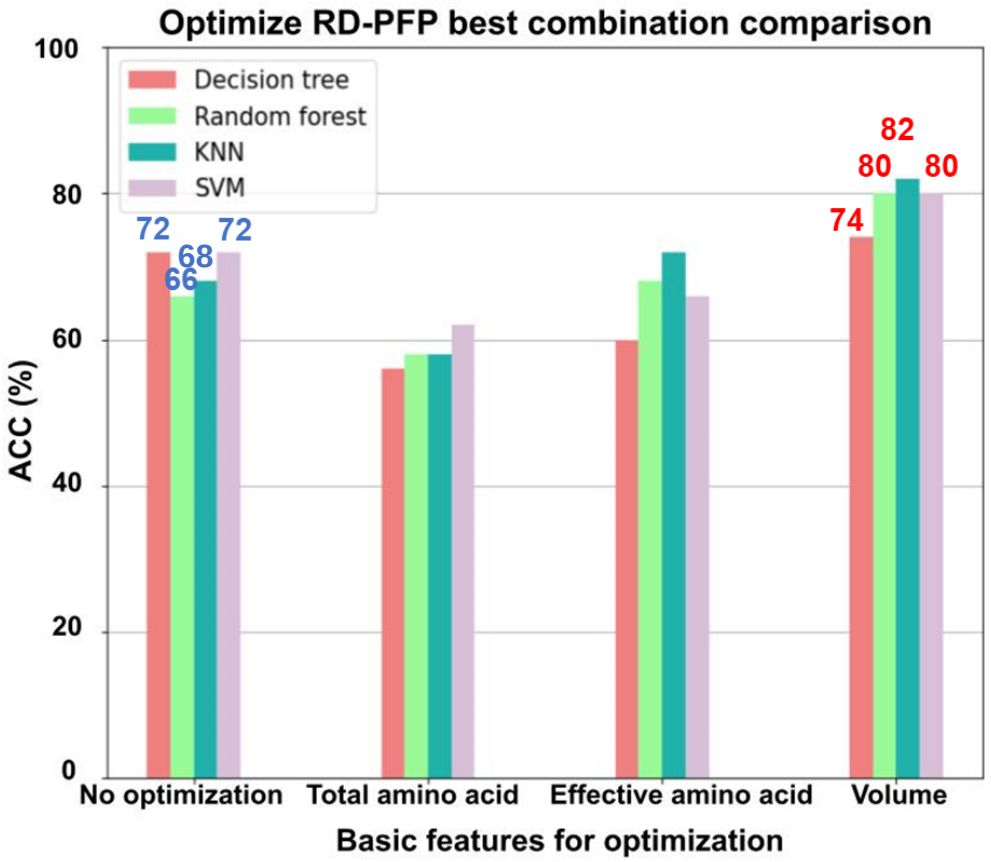
Features for optimization	Total amino acid number			Effective amino acid number			Volume		
Metrics / Classifier	Precision	Recall	F_1 *score*	Precision	Recall	F_1 *score*	Precision	Recall	F_1 *score*
DT	0.50	0.48	0.49	0.52	0.55	0.53	**0.75**	**0.74**	**0.74**
RF	0.64	0.55	0.59	0.67	0.65	0.66	**0.83**	**0.80**	**0.81**
KNN	0.51	0.50	0.50	0.62	0.62	0.62	**0.84**	**0.82**	**0.83**
SVM	0.52	0.51	0.51	0.67	0.63	0.65	**0.80**	**0.79**	**0.80**

Experiments and Results

Fingerprint Generation

International Networks and High-performance Computation Lab

- **RD-PFP Optimization**
- **Effective amino acid number more accurately represent the number of amino acids.**
- **Ideal RD-PFP**
  - Optimize feature: **Volume**



Experiments and Results

Ideal RD-PFP and Basic Features Performance

International Networks and High-performance Computation Lab

- In order to allow the classifier **to obtain more information** about DNA mimic proteins and improve the classification effectiveness

- Input the **basic protein features** together with **Ideal RD-PFP**

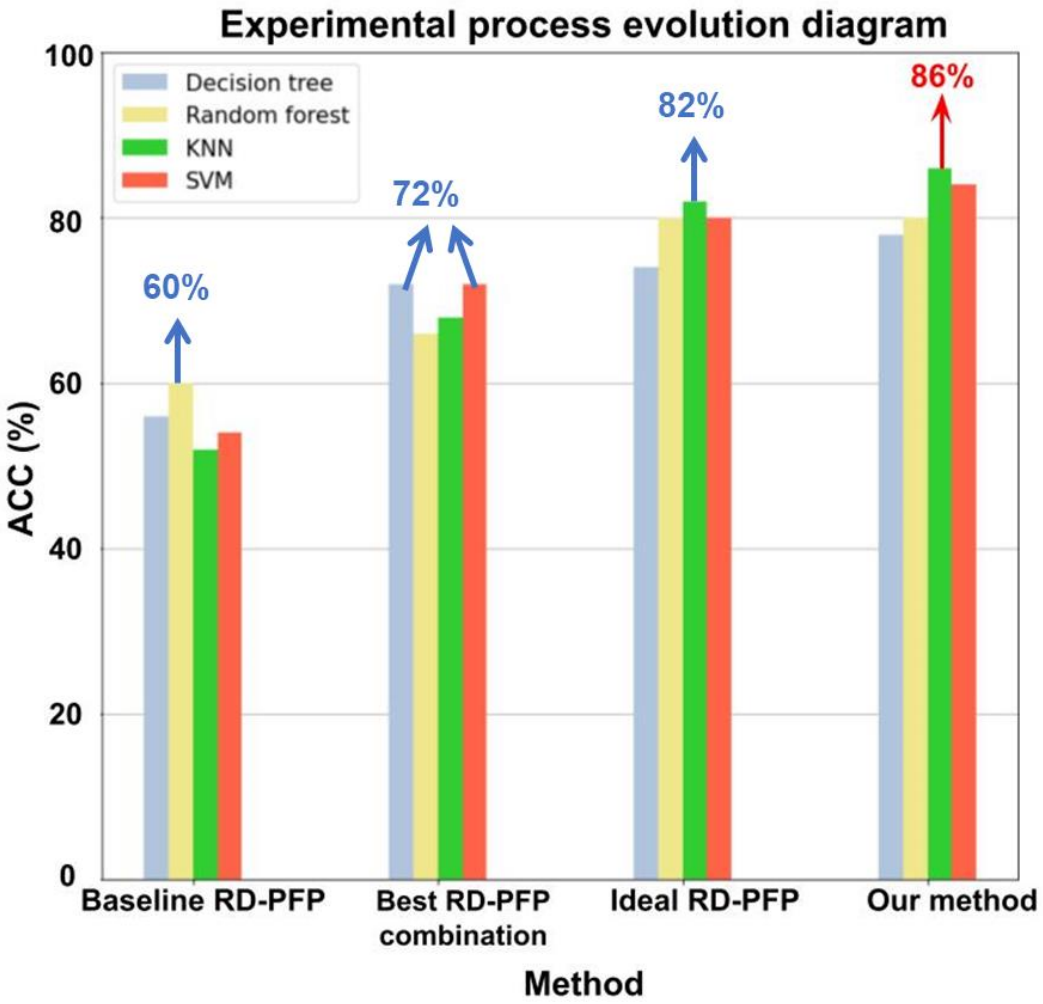
Metrics / Classifier	*Precision*	*Recall*	*F_1 score*	MCC	*ACC (%)*
DT	0.81	0.80	0.80	0.60	78
RF	0.89	0.88	0.89	0.77	80
KNN	**0.93**	**0.88**	**0.90**	**0.80**	**86**
SVM	0.86	0.80	0.83	0.65	84

Experiments and Results

Ideal RD-PFP and Basic Features Performance

International Networks and High-performance Computation Lab

- **Utilize KNN as the final classifier for the model.**

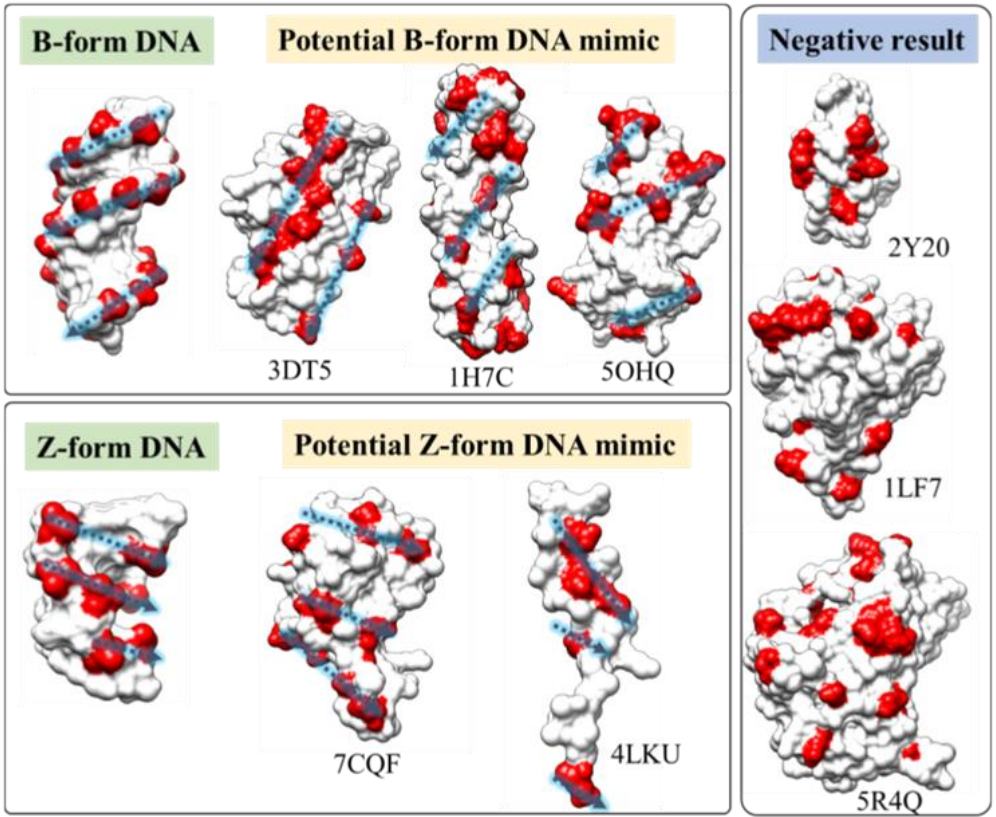


Experiments and Results

Protein Prediction

International Networks and High-performance Computation Lab

- Before prediction, **obtain the basic features and fingerprint features of these proteins (Ideal RD-PFP).**
- **3DT5, 1H7C, and 5OHQ**, **right-handed helical** in the surface negative charge distribution, similar to **B-form DNA**.
- **4LKU** and **7CQF**, **left-hand** surface negative charge distribution similar to **Z-form DNA**.
- **Non-DNA mimics** show more dispersed and disordered distribution of negatively charged amino acids.

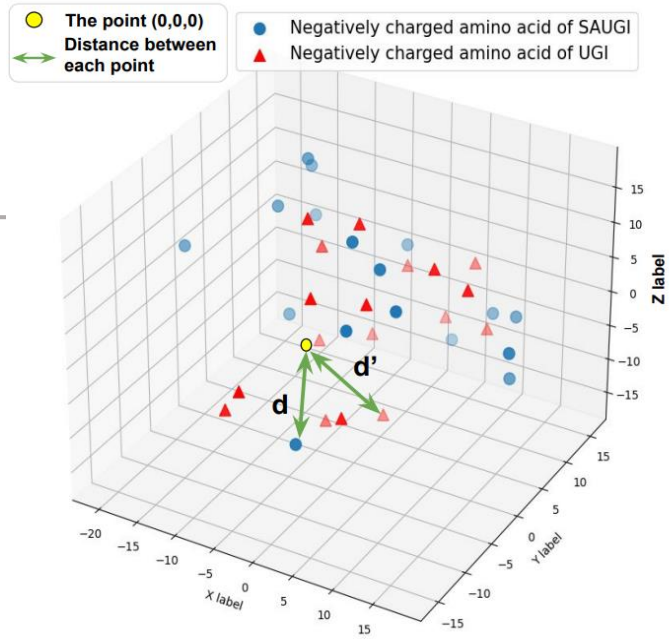


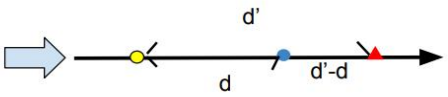
Discussion

- Mean Square Error
- Maximum Bipartite Matching
- Method Compare
- Further Application of RD-PFP

Discussion

Mean Square Error

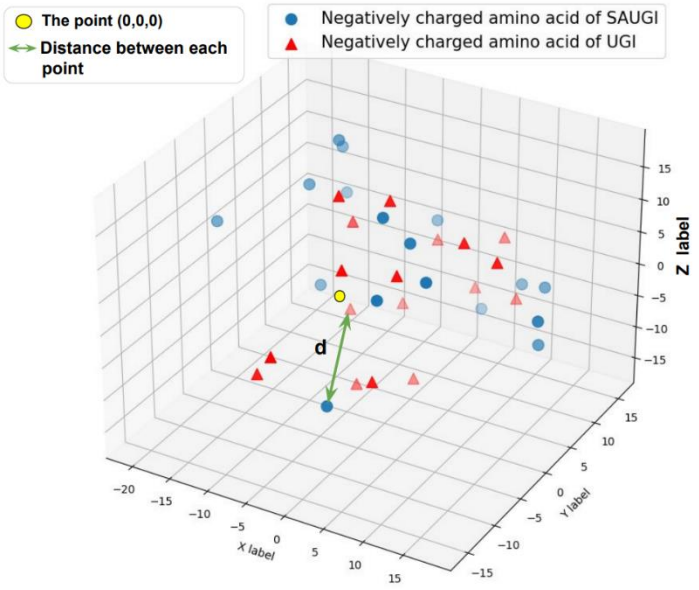




- **Function:**
  - $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
  - n: sample size
  - y_i : the actual data value
  - $\hat{y}_i$: the predicted data value
- **One-dimensional method**
- Transform three-dimensional coordinates to a single coordinate.
- **Sum the squares of the smallest distance differences for each atom and divide it by the number of negatively charged amino acids** in the protein with unknown properties.
- Protein with unknown properties is assumed to possess the same properties as the protein with known properties that has the smallest MSE.

Discussion

Mean Square Error



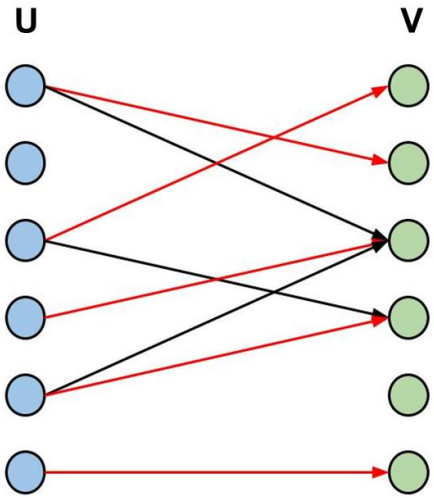
- **Three-dimensional method**

- **Calculate the distance difference in three-dimensional space** between a protein with unknown properties and a protein with known properties.
- **Sum the squares of the smallest distance differences** for each amino acids and **divide it by the number of negatively charged amino acids** in the protein with unknown properties.

- Protein with unknown properties have the same properties as the protein with known properties that has the smallest MSE.

Discussion

Maximum Bipartite Matching



- **Bipartite graph** consists of two mutually exclusive and independent sets**, U and V.**

- All edges connect a vertex from set U to a vertex from set V.

- **Bipartite matching**: selecting a set of edges in a graph no two edges share an endpoint.

- **Maximum matching** is achieved when the maximum number of edges is chosen.

Discussion

Maximum Bipartite Matching

- Represent **negatively charged atoms** as vertices in a bipartite graph

- **One set** (U): the collection of negatively charged amino acids for proteins with **known properties**,
- **Other set** (V): the collection of negatively charged amino acids for proteins with **unknown properties**.

- **Establishing edges** between different sets of negatively charged atoms based on a **distance threshold.**

- The **maximum matching identifies protein pairs** suggesting that they **share the same properties**.

Discussion

Maximum Bipartite Matching

- Employ both one-dimensional and three-dimensional methods.

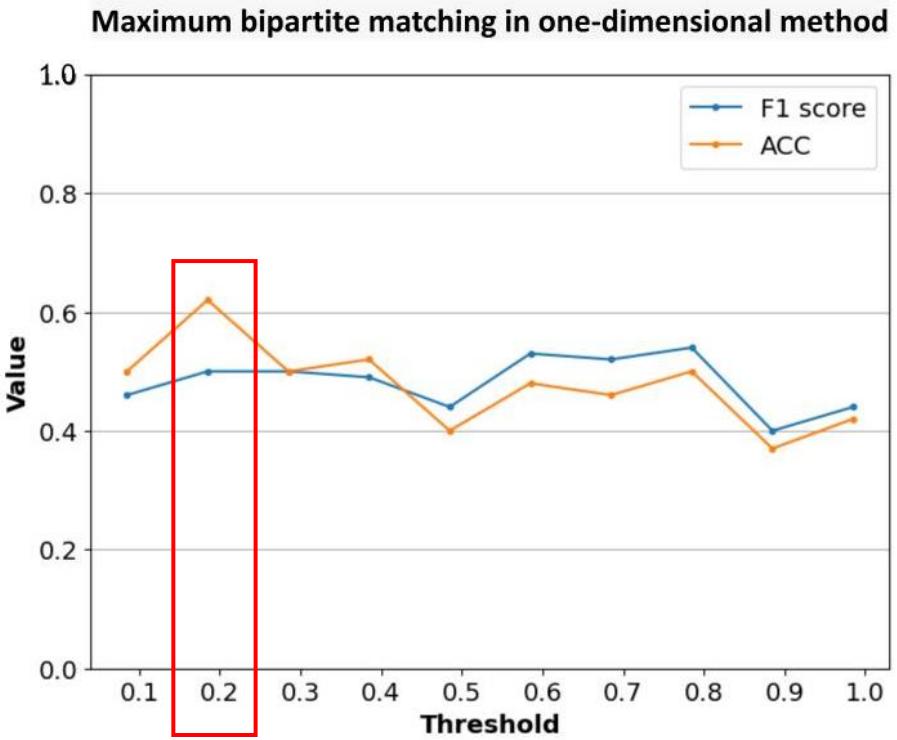
- **One-dimensional method**: calculate the distance value for the distance between each amino acid by **computing the difference between its distance and the distance to the point (0, 0, 0).**

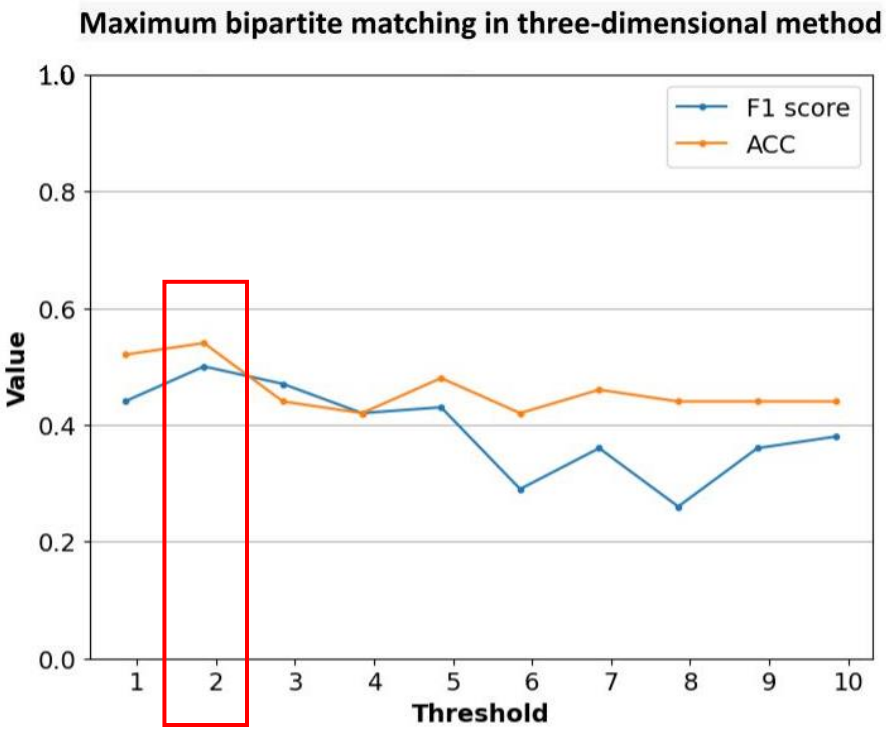
- **Three-dimensional method**: calculate **the distance difference between two amino acids** in three-dimensional space as the similarity value for their interatomic distance.

- If the **distance value between two amino acids is less than the threshold**, it indicates that their spatial distribution is similar. In such cases, we **establish an edge between the two different groups of atoms**.

Discussion

Method Compare





Discussion
Method Compare

International Networks and High-performance Computation Lab

Metrics / Method	*Precision*	*Recall*	*F_1 score*	MCC	*ACC (%)*
MSE one-dimensional	0.43	0.68	0.53	-0.03	0.56
MSE three-dimensional	0.28	0.26	0.27	-0.18	0.48
MBP one-dimensional	0.48	0.53	0.50	-0.01	0.62
MBP three-dimensional	0.41	0.63	0.50	-0.05	0.54
Our method	**0.93**	**0.88**	**0.90**	**0.80**	**0.86**

Discussion

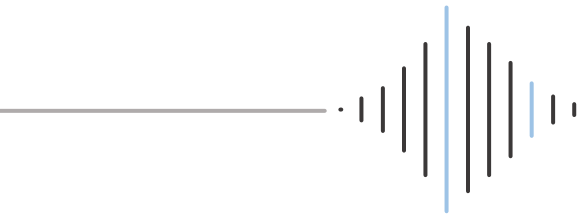
Further Application of RD-PFP

- Beyond the prediction of DNA mimic proteins.

- Utilized to calculate fingerprints of proteins before and after the evolution of disulfide bonds, aiding in the **identification of proteins with potential evolutionary conservation.**

- Employed to identify fingerprints of DNA-binding proteins, **facilitating the assessment of their ability to specifically bind to certain DNA or RNA sequences,** based on their unique positively charged fingerprints.

- These applications demonstrate the versatility and utility of the RD-PFP algorithms in protein fingerprinting and analysis.



Conclusion

Conclusion

- Our study represents a **pioneering** application of machine learning to predict **DNA mimic proteins**.
- Proposed the **RD-PFP algorithm** to extract protein fingerprints.
- Presented **several potential DNA mimic proteins** and non-DNA mimic proteins.
- Proposed further applications of the RD-PFP algorithm.

Thanks for listening