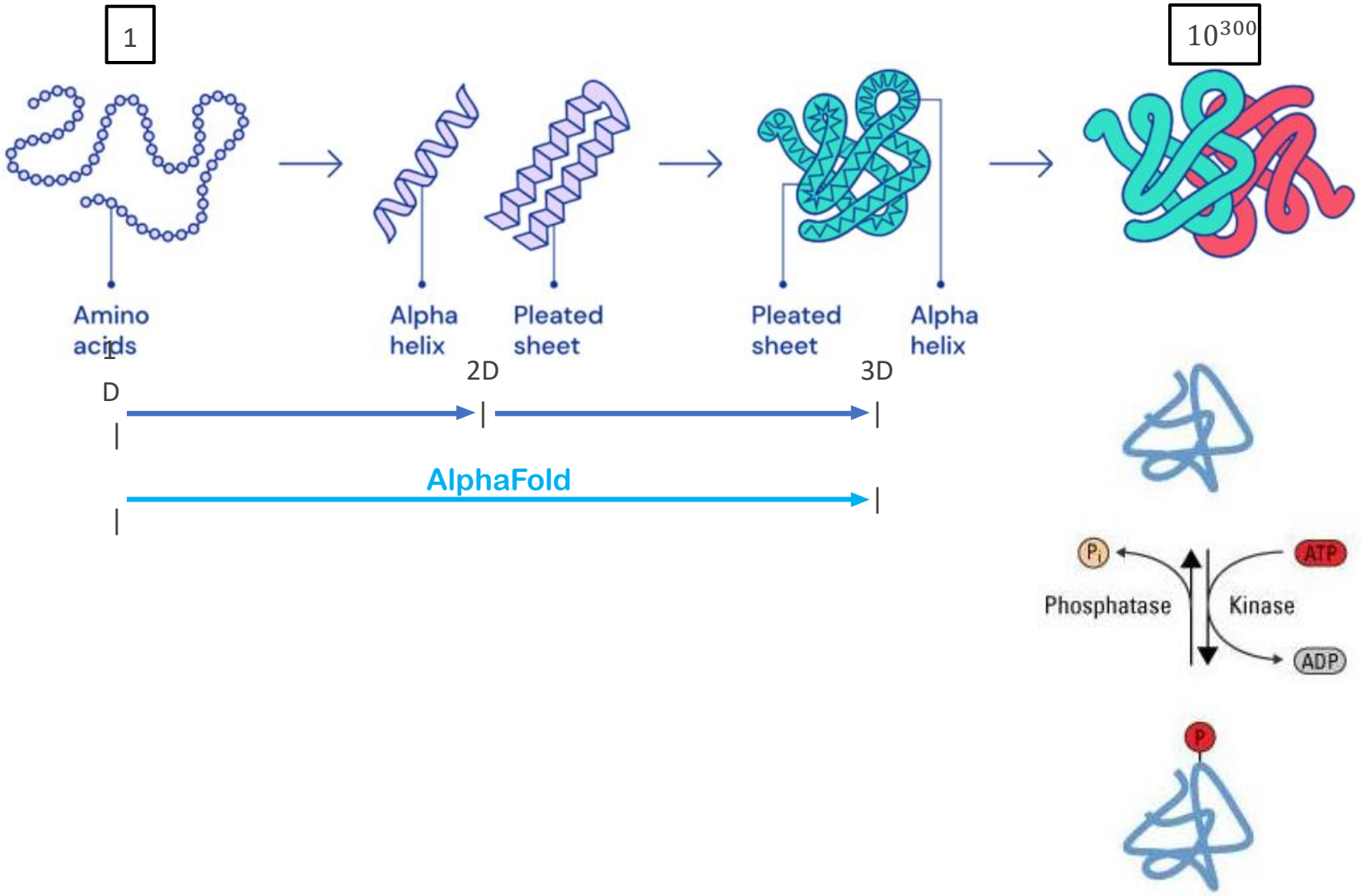
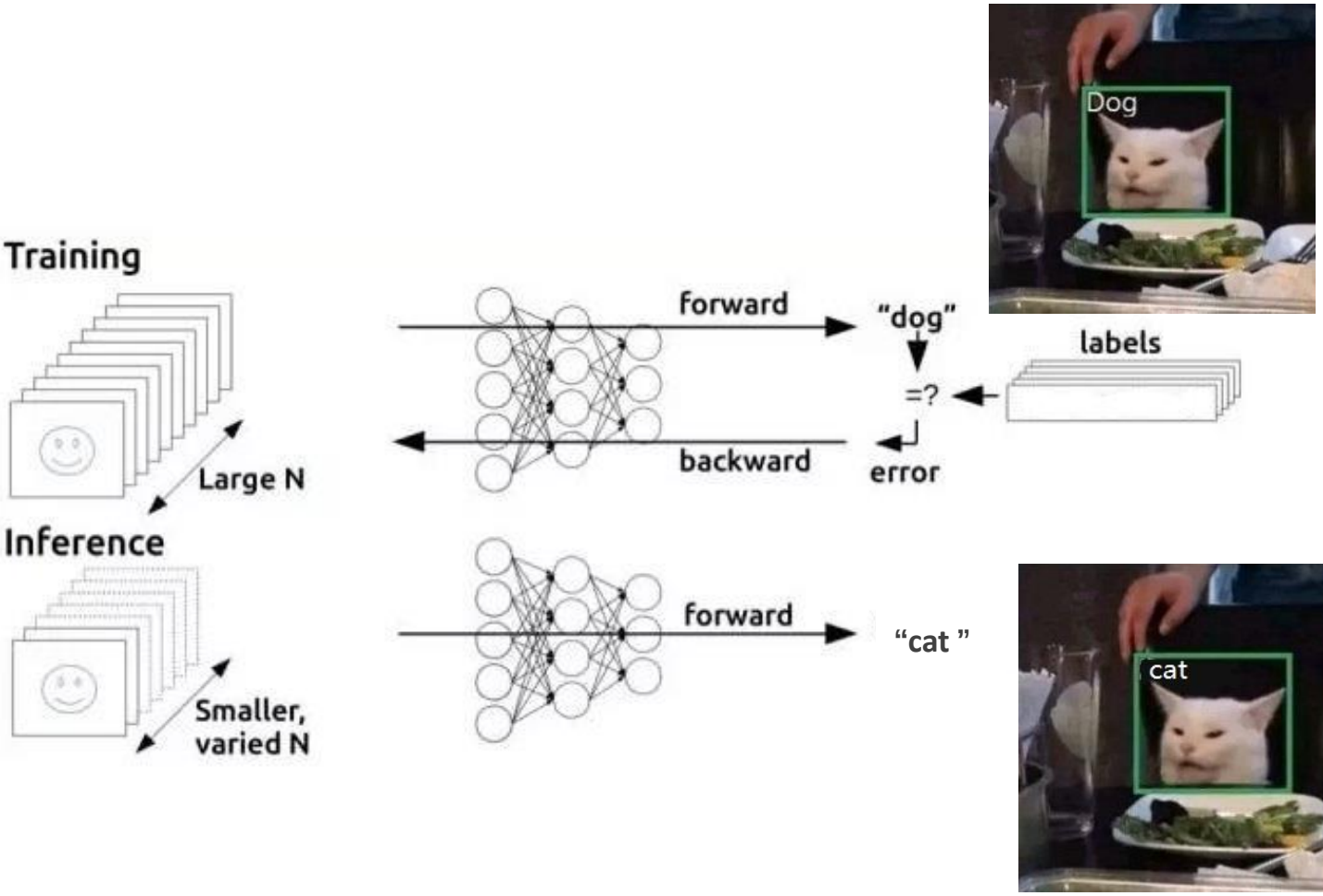
Work Report

贾宁欣

06-06-2021

Protein Folding




Training
Large N
Inference
Smaller,
varied N
forward
backward
"dog"
=?
error
labels
forward
"cat "
Dog
cat

Nucleic Acids Research, 2021 **1**
doi: 10.1093/nar/gkab044

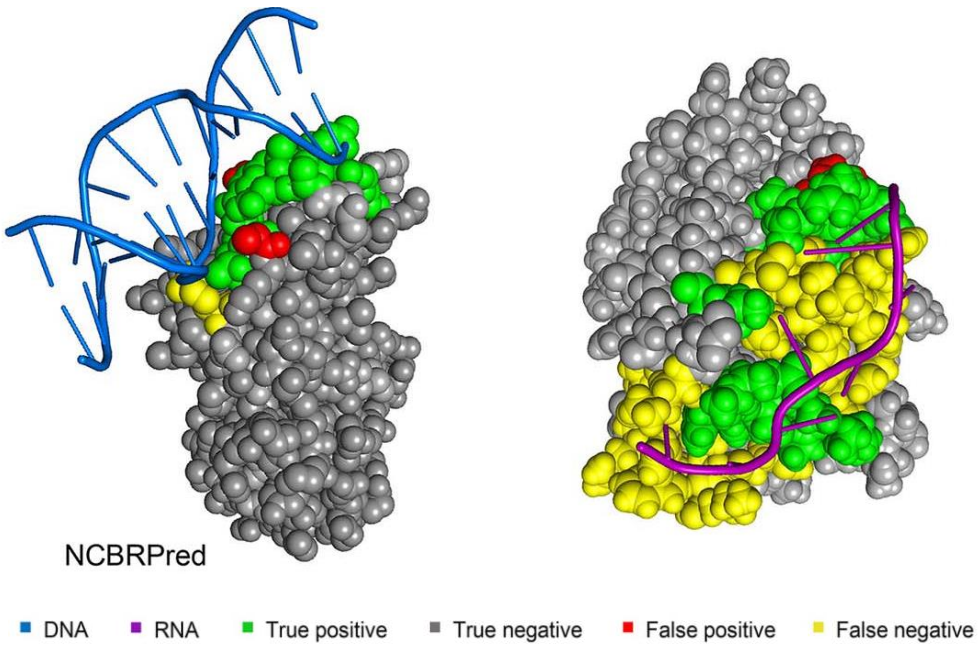
GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues

Ying Xia[1], Chun-Qiu Xia[1], Xiaoyong Pan[1,*] and Hong-Bin Shen[1,2,*]

[1]Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China and [2]School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China

Received December 16, 2020; Editorial Decision January 09, 2021; Accepted February 09, 2021

NCBRPred



Problem:

cross-prediction

(If a DNA-binding residue predictor is only trained with DNA-binding proteins and does not consider the RNA-binding proteins, it can accurately predict the DNA-binding residues but also prefers to identify the RNA-binding residues as DNA-binding residues.)

Reason:

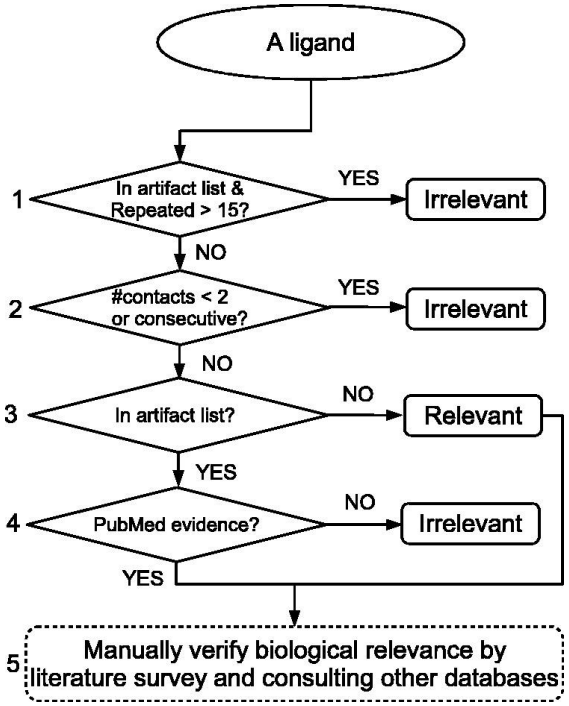
DNA-binding residues and RNA-binding residues share some similar characteristics.

Benchmark Datasets

D1096–D1103 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 18 October 2012*
doi:10.1093/nar/gks966

BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions

Jianyi Yang, Ambrish Roy and Yang Zhang*



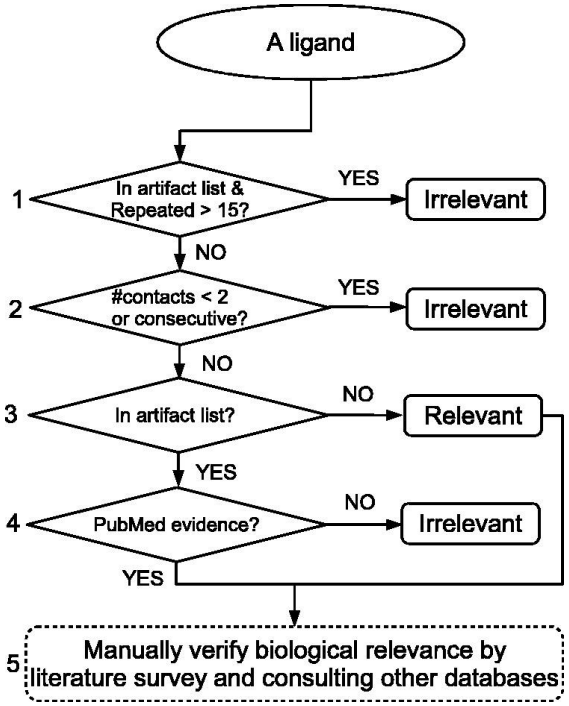
		number of …	DNA	RNA	DNA-RNA
BioLiP database		binding proteins	4344	1558	440
train datasets bef 2016.01	benchmark datasets		573	495	0
	without dada augmentation	binding residues	11074	11756	
		non-binding residues	148809	125143	
		bi/non-bi ratio	0.074	0.094	
	with dada augmentation	binding residues	14479	14609	
		non-binding residues	145404	122290	
		bi/non-bi ratio	0.100	0.119	
		increase	30.7%	19.5%	

Benchmark Datasets

D1096–D1103 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 18 October 2012*
doi:10.1093/nar/gks966

BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions

Jianyi Yang, Ambrish Roy and Yang Zhang*



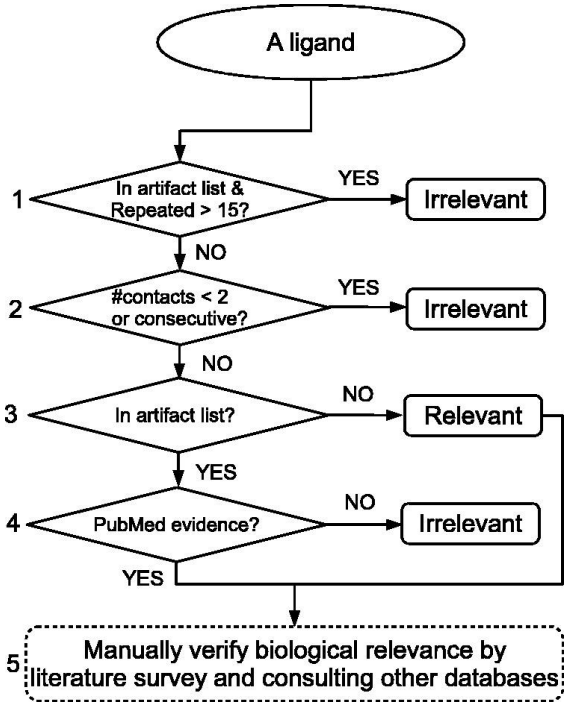
		number of …	DNA	RNA	DNA-RNA
BioLiP database		binding proteins	4344	1558	440
train datasets bef 2016.01	benchmark datasets		573	495	0
	without dada augmentation	binding residues	11074	11756	
		non-binding residues	148809	125143	
		bi/non-bi ratio	0.074	0.094	
	with dada augmentation	binding residues	14479	14609	
		non-binding residues	145404	122290	
		bi/non-bi ratio	0.100	0.119	
		increase	30.7%	19.5%	

Benchmark Datasets

D1096–D1103 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 18 October 2012*
doi:10.1093/nar/gks966

BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions

Jianyi Yang, Ambrish Roy and Yang Zhang*



		number of …	DNA	RNA	DNA-RNA
BioLiP database		binding proteins	4344	1558	440
train datasets bef 2016.01	benchmark datasets		573	495	0
	without dada augmentation	binding residues	11074	11756	
		non-binding residues	148809	125143	
		bi/non-bi ratio	0.074	0.094	
	with dada augmentation	binding residues	14479	14609	
		non-binding residues	145404	122290	
		bi/non-bi ratio	0.100	0.119	
		increase	30.7%	19.5%	

Data Augmentation

W20–W25 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh435

BLAST: at the core of a powerful and diverse set of sequence analysis tools

Scott McGinnis* and Thomas L. Madden

asses the sequence identity between protein chain pairs

>0.8

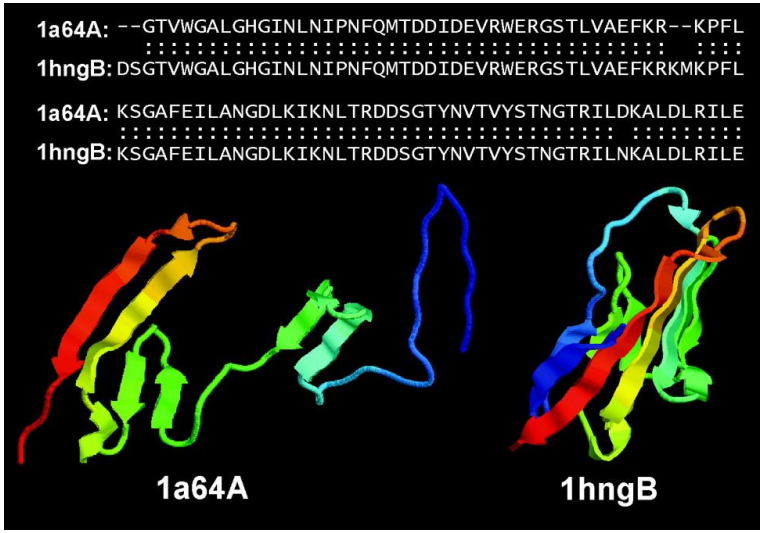
2302–2309 *Nucleic Acids Research, 2005, Vol. 33, No. 7*
doi:10.1093/nar/gki524

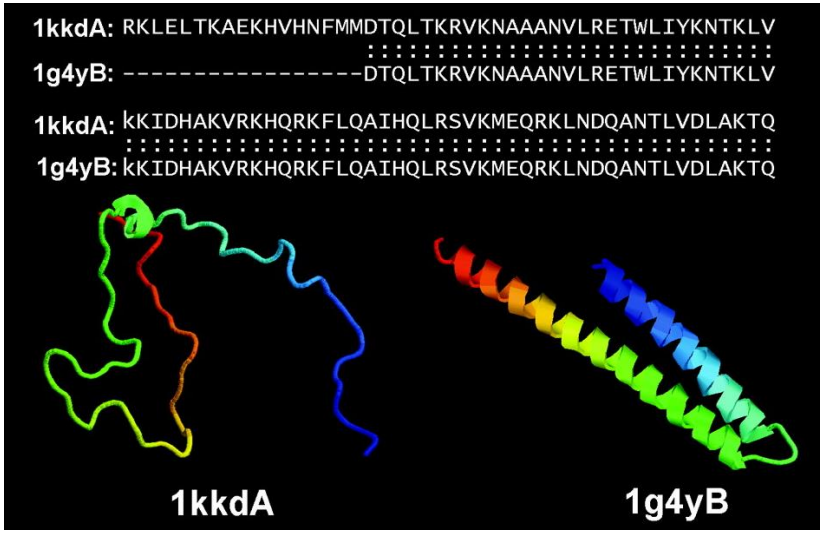
TM-align: a protein structure alignment algorithm based on the TM-score

Yang Zhang and Jeffrey Skolnick*

asses the structural similarity between protein chain pairs

TM scores >0.5





Data Augmentation

W20–W25 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh435

BLAST: at the core of a powerful and diverse set of sequence analysis tools

Scott McGinnis* and Thomas L. Madden

asses the sequence identity between protein chain pairs

>0.8

2302–2309 *Nucleic Acids Research, 2005, Vol. 33, No. 7*
doi:10.1093/nar/gki524

TM-align: a protein structure alignment algorithm based on the TM-score

Yang Zhang and Jeffrey Skolnick*

asses the structural similarity between protein chain pairs

TM scores >0.5

transferring binding annotations

transfer annotations of protein chains in the same cluster into the longest chain

CD-HIT (30%)

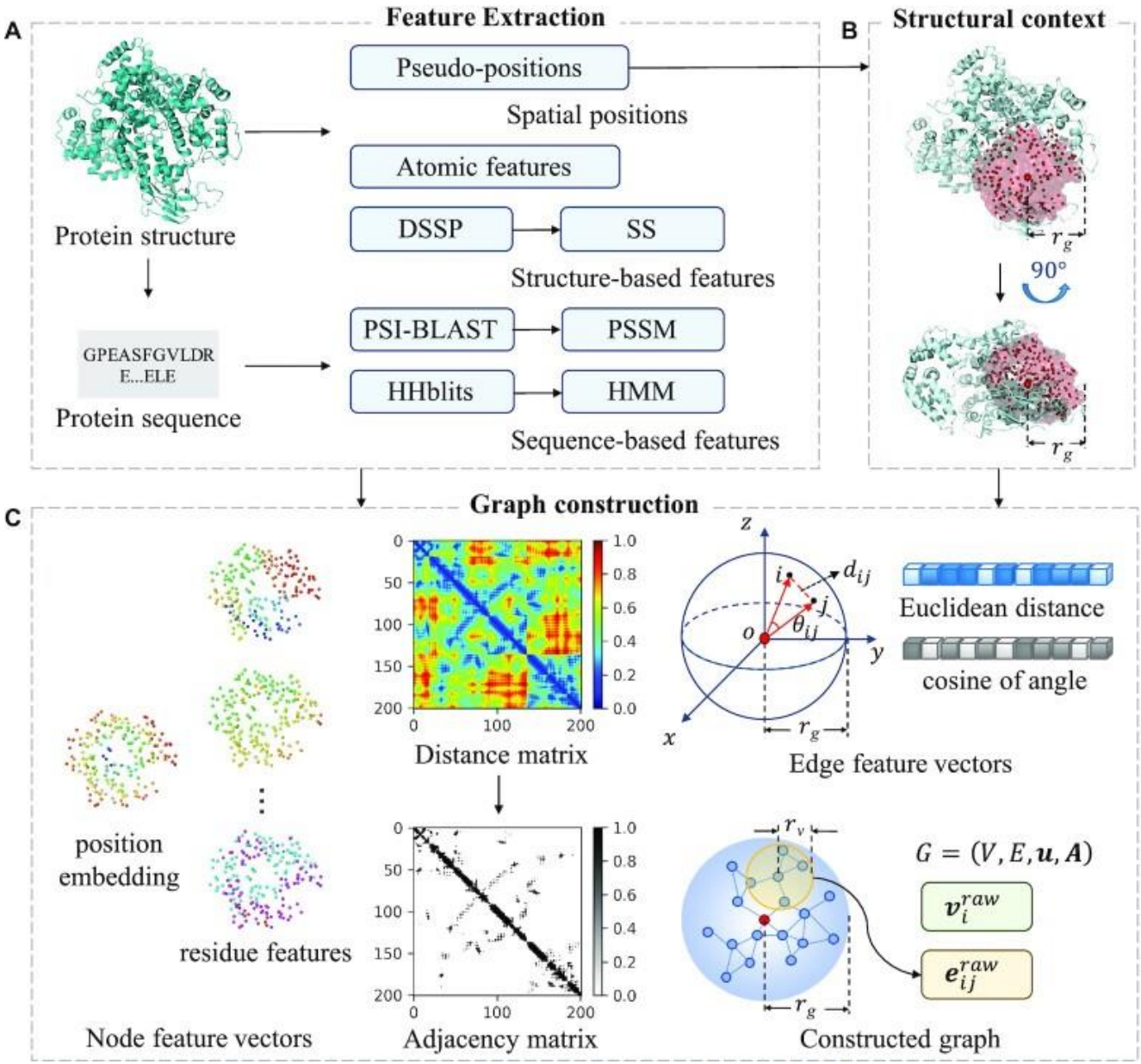
remove the redundant protein chains

benchmark datasets

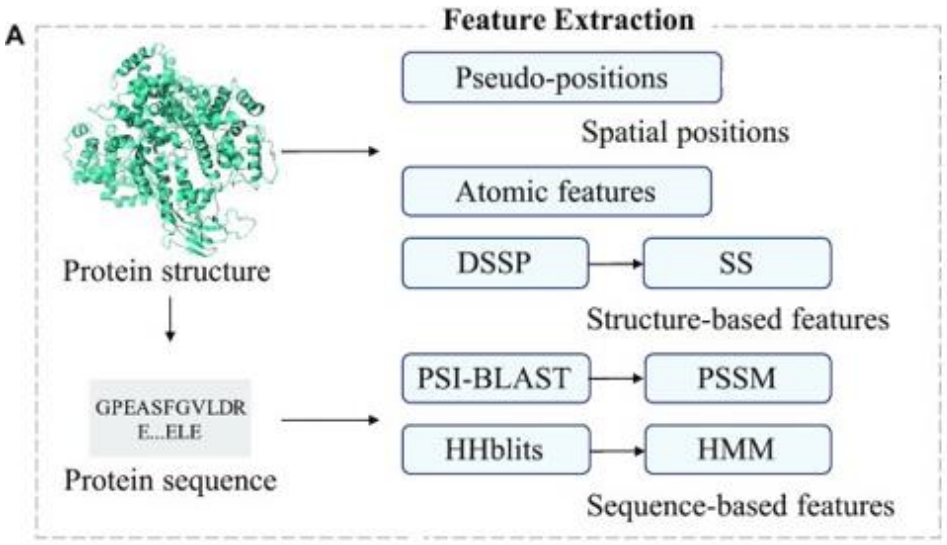
Type	Dataset	$N_{protein}$[a]	N_{pos}[b]	N_{neg}[c]	PNratio[d]
DNA	DNA-573_Train	573	14479	145404	0.100
	DNA-129_Test	129	2240	35275	0.064
RNA	RNA-495_Train	495	14609	122290	0.119
	RNA-117_Test	117	2031	35314	0.058

[a] Number of proteins.
[b] Number of binding residues.
[c] Number of non-binding residues.
[d] PNratio = N_{pos}/N_{neg}.

Pipeline of Graph Construction

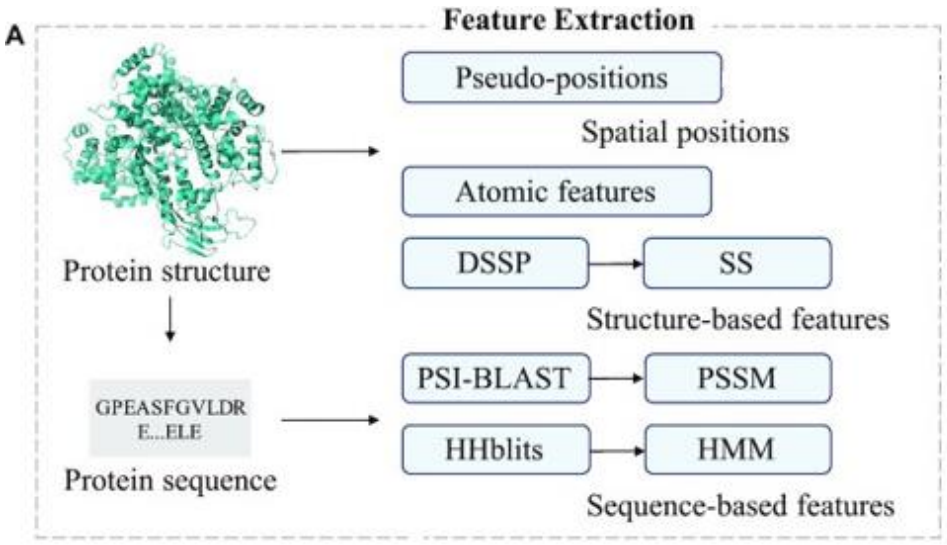


Feature (residue-level) extraction



Pseudo-positions = centroid of a residue including both backbone and side-chain atoms of the residue **L x 3**

Feature (residue-level) extraction

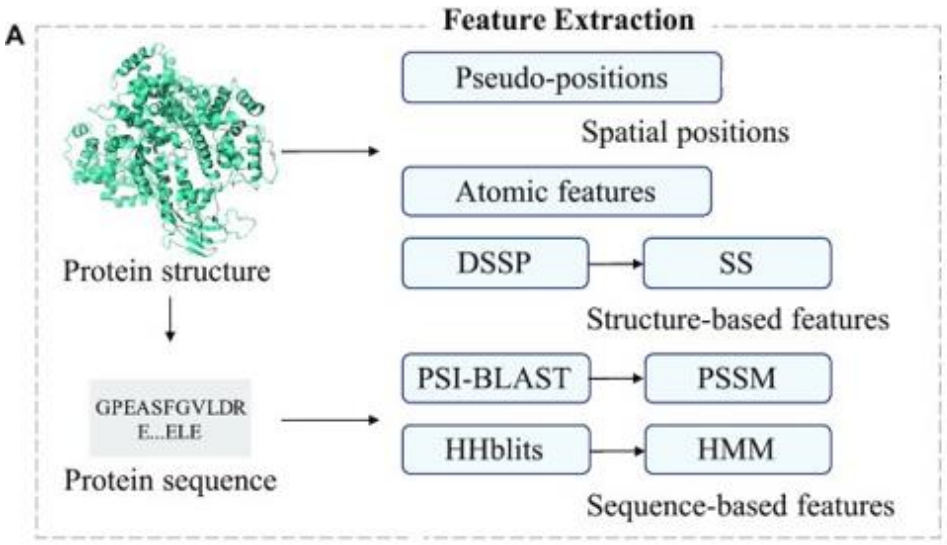


Pseudo-positions = centroid of a residue including both backbone and side-chain atoms of the residue **L x 3**

Atomic features = average the sth feature of all the atoms as the sth atomic feature x_s of the residue **L x 7**

$$x_s = \frac{1}{N_a}\left(\sum_{t=1}^{t=N_a} f_{s,t}\right) \quad \{x_s\}_{s=1,\ldots,7}$$

Feature (residue-level) extraction



Pseudo-positions = centroid of a residue including both backbone and side-chain atoms of the residue **L x 3**

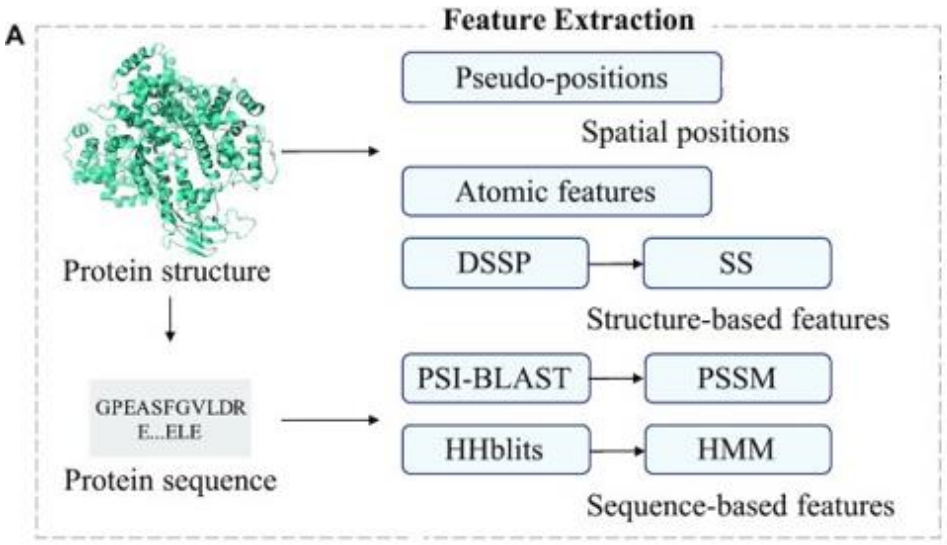
Atomic features = average the sth feature of all the atoms as the sth atomic feature x_s of the residue **L x 7**

SS = secondary structure profile : residue water-exposed surface X 1 **L x 14**

bond and torsion angles X 5

one-hot encoded secondary structure with eight states X8

Feature (residue-level) extraction



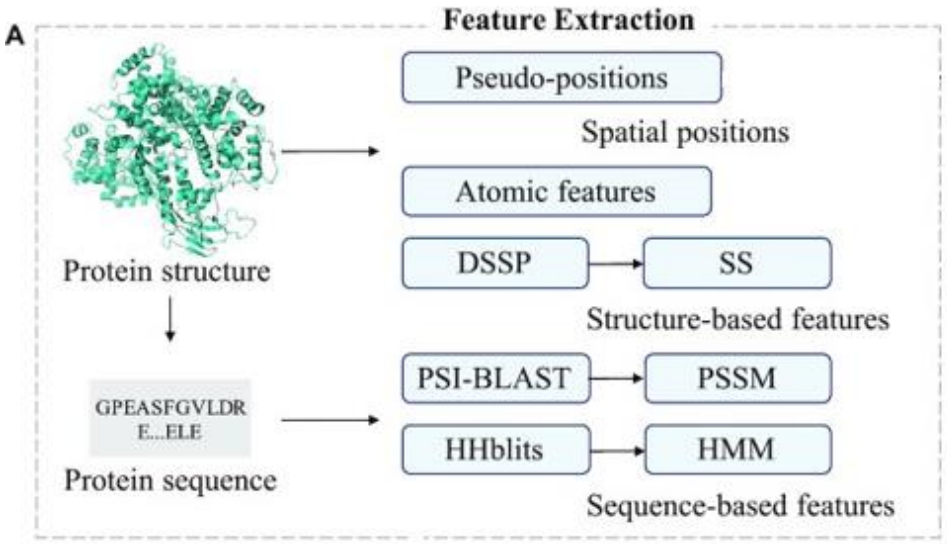
Pseudo-positions = centroid of a residue including both backbone and side-chain atoms of the residue **L x 3**

Atomic features = average the sth feature of all the atoms as the sth atomic feature x_s of the residue **L x 7**

SS = secondary structure profile **L x 14**

PSSM = position-specific scoring matrix (is normalized to the range [0, 1]) **L x 20**

Feature (residue-level) extraction



Pseudo-positions = centroid of a residue including both backbone and side-chain atoms of the residue **L x 3**

Atomic features = average the sth feature of all the atoms as the sth atomic feature x_s of the residue **L x 7**

SS = secondary structure profile **L x 14**

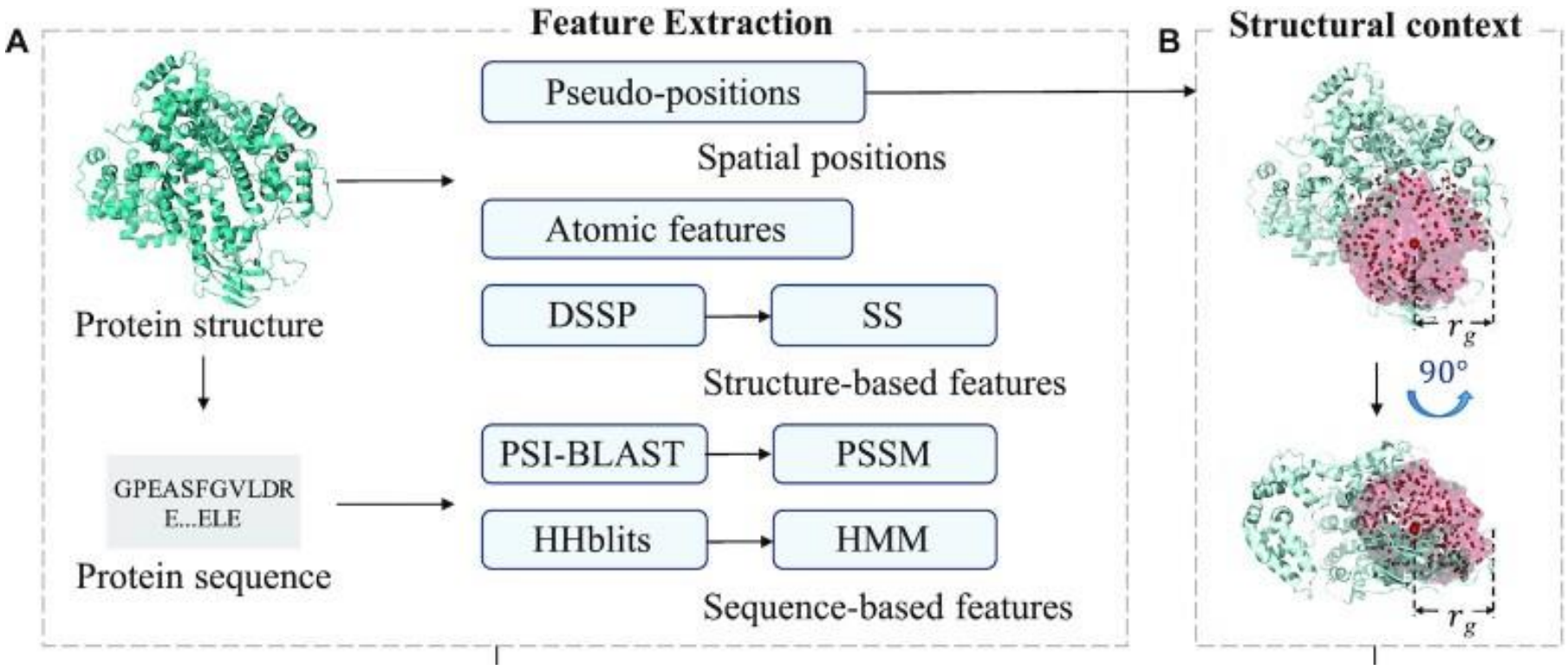
PSSM = position-specific scoring matrix (is normalized to the range [0, 1]) **L x 20**

HMM = hidden Markov matrix : observed frequencies for 20 amino acids in homologous sequences X 20 **L x 30**

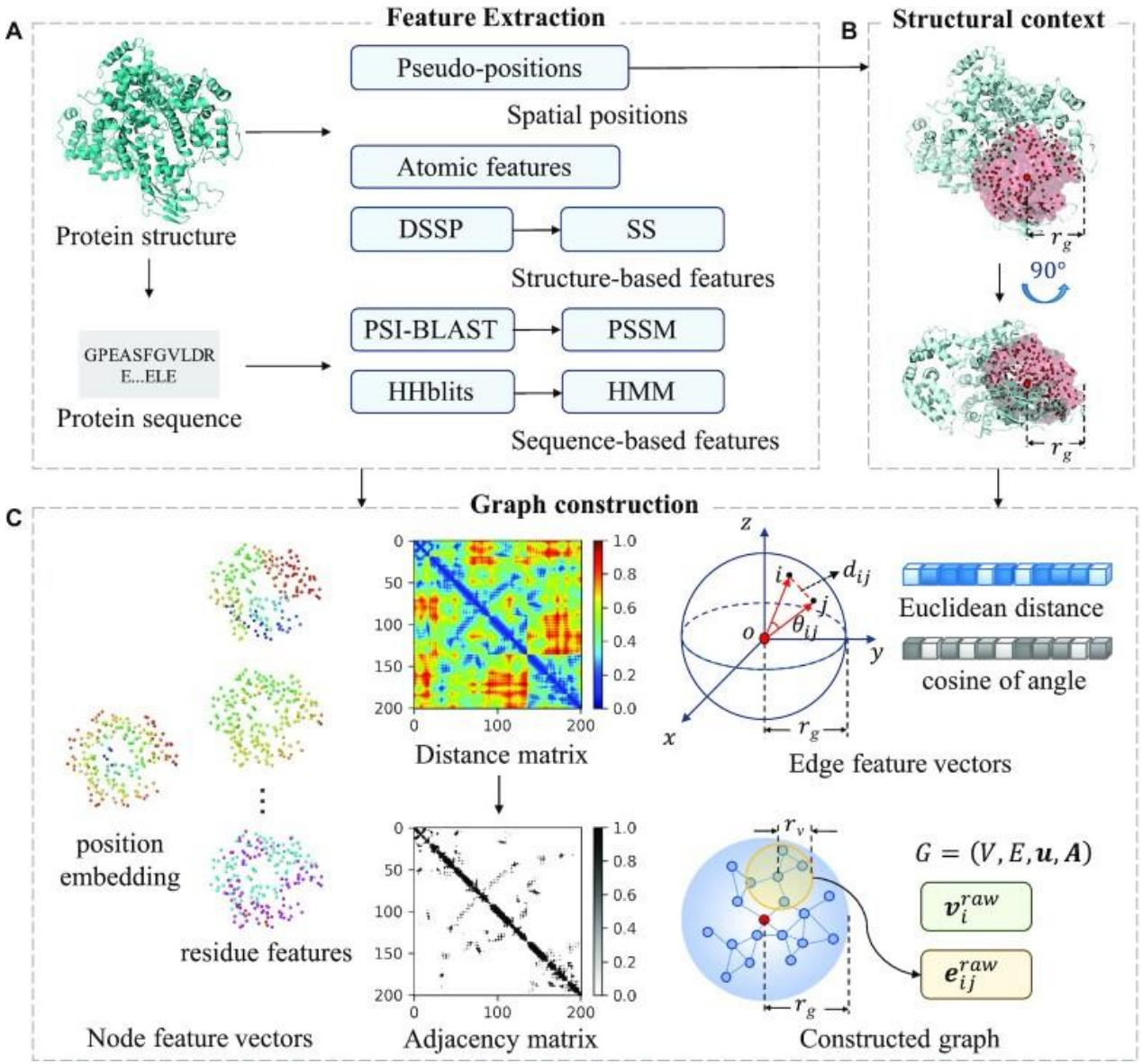
transition frequencies X 7

local diversities X3

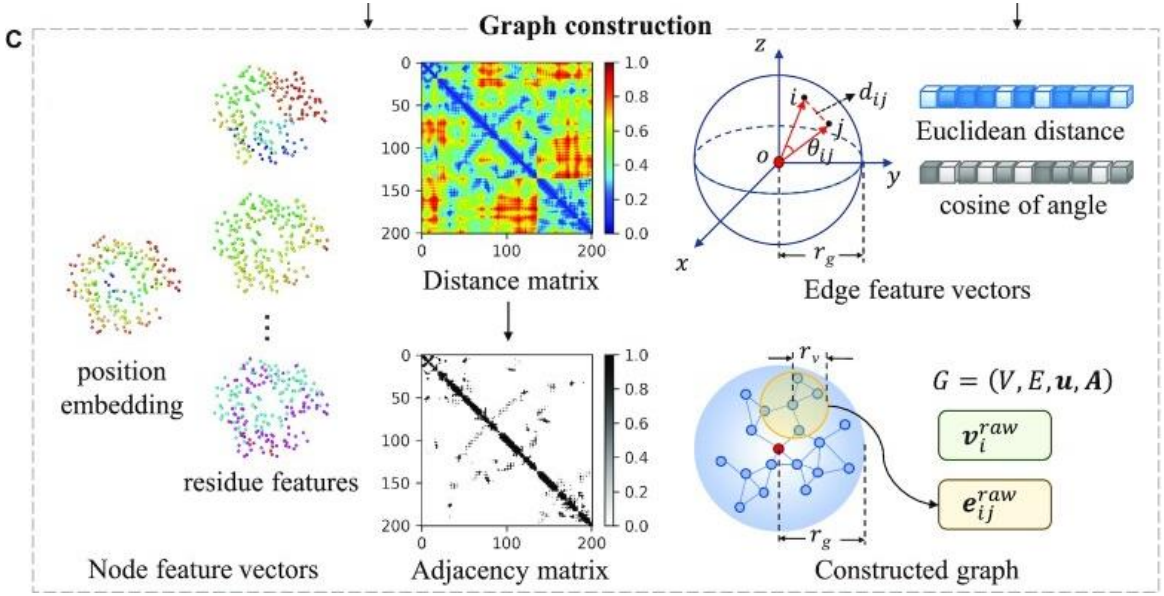
Structural Context Extraction



Graph Construction



Graph Construction



$$G = (V, E, u, A)$$

$$V = \{v_i\}_{i=1,\dots,N_v}$$

Raw feature vector :

the concatenation of the **position embedding** and the **residue features** of node .

$$PE_i = \frac{1}{r_g}|\overrightarrow{p_0 p_i}|$$

Distance matrix :

calculated based on **pseudo positions** of residues.

Adjacency matrix :

$$A_{\mathrm{ij}_{(N_v \times N_v)}} = \begin{cases} 1, & if\, D_{ij} < r_v \\ 0, & if\, D_{ij} \geq r_v \end{cases} \qquad D_{ij} = |\overrightarrow{p_i p_j}|$$

apply the **binary threshold** r_v on the distance matrix, which records the connections of nodes.

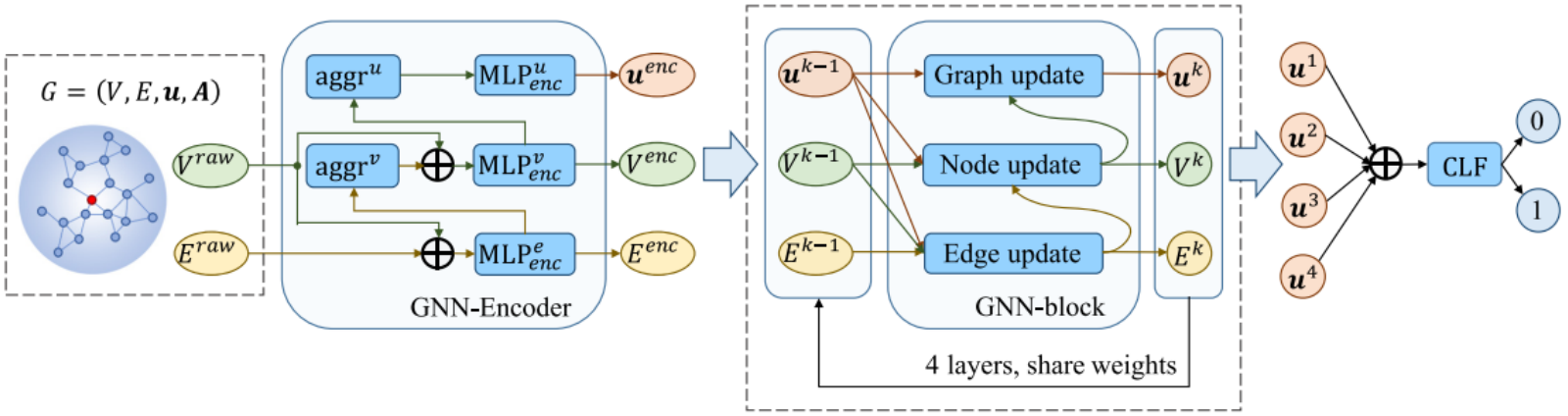
Edge feature vectors :

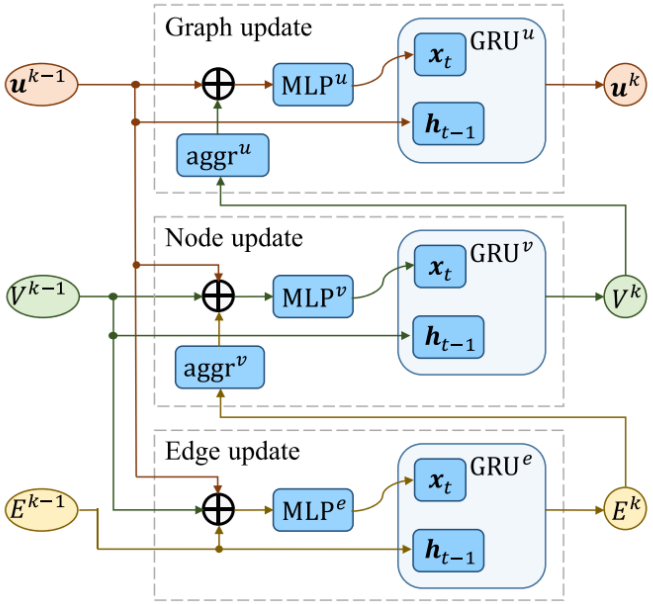
e_{ij} : the **Euclidean distance** between the two adjacent nodes.

$$E = \{e_{\mathrm{ij}} | A_{\mathrm{ij}} = 1\} \quad cos(\theta_{ij}) = \frac{\overrightarrow{p_0 p_i} \cdot \overrightarrow{p_0 p_j}}{|\overrightarrow{p_0 p_i}||\overrightarrow{p_0 p_j}|}$$

θ_{ij} : the **cosine of the angle** between the two vectors from the sphere center to the two adjacent nodes.

Hierarchical Graph Neural Networks (HGNN)





ATP Binding Residues Prediction

Train	protein	388
	residues	147743
	positive examples	5657
	negative examples	142086

Test	protein	41
	residues	14833
	site	674
	others	14159

Ligand	Method	REC	PRE	MCC	AUC
ATP[e] (ATP-41_Test)	COACH	0.632	0.703	0.652	N/A
	ATPBind	0.631	0.756	0.677	0.915
	TargetS	0.516	0.689	0.580	N/A
	S-SITE	0.570	0.505	0.513	0.801
	DELIA	**0.642**	**0.758**	**0.685**	**0.947**

features	Thres	Sen (%)	Spe (%)	Acc (%)	Pre (%)	MCC	F1-sco	AUC
TMSITE+ILBR	0.28	63.95	98.96	97.37	74.5	0.677	0.688	0.884
	0.31	58.16	99.47	97.60	84.12	0.688	0.697	0.905
	0.34	66.02	99.17	97.67	79.18	0.711	0.702	0.932
	0.34	66.02	99.17	97.67	79.18	0.711	0.720	0.932
	0.30	68.10	98.86	97.56	75.74	0.706	0.717	0.933
	0.36	67.21	99.05	97.60	77.04	0.707	0.718	0.936
	0.30	66.47	99.12	97.64	78.32	0.709	0.719	0.935

实验结果不能用MaxMCC

```
(base) [junh@csbio lstm]$ java -jar MaxMCCReporter.jar /data0/junh/stu/ningxinJ/sy/lstm
/lstm_lib/lstm_ti_1000-100.prod
Ind      Thres   Sen(%)   Spe(%)   Acc(%)   Pre(%)    MCC    F1-sco    AUC
  1       0.30    66.47    99.12    97.64    78.32   0.709     0.719  0.935
```