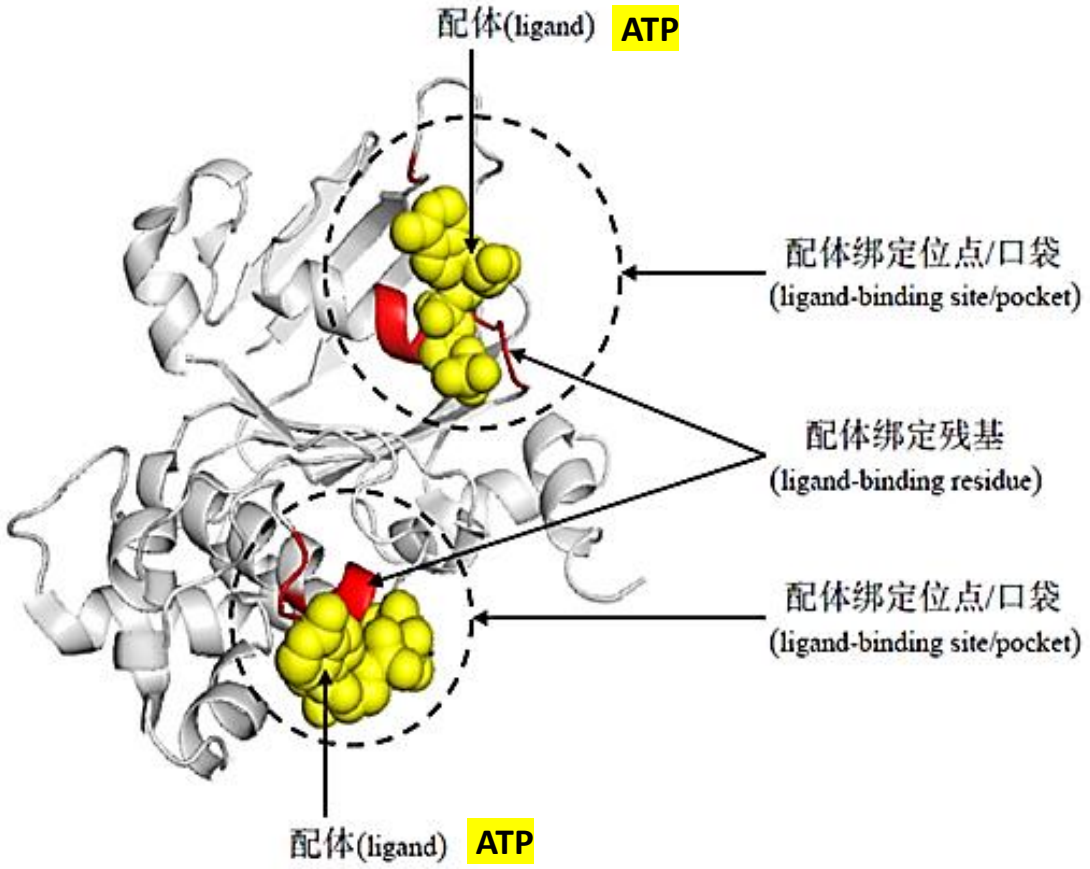
Work Report

贾宁欣

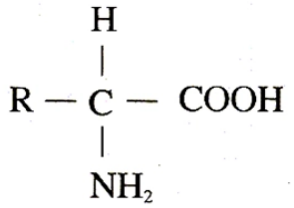
11-27-2021

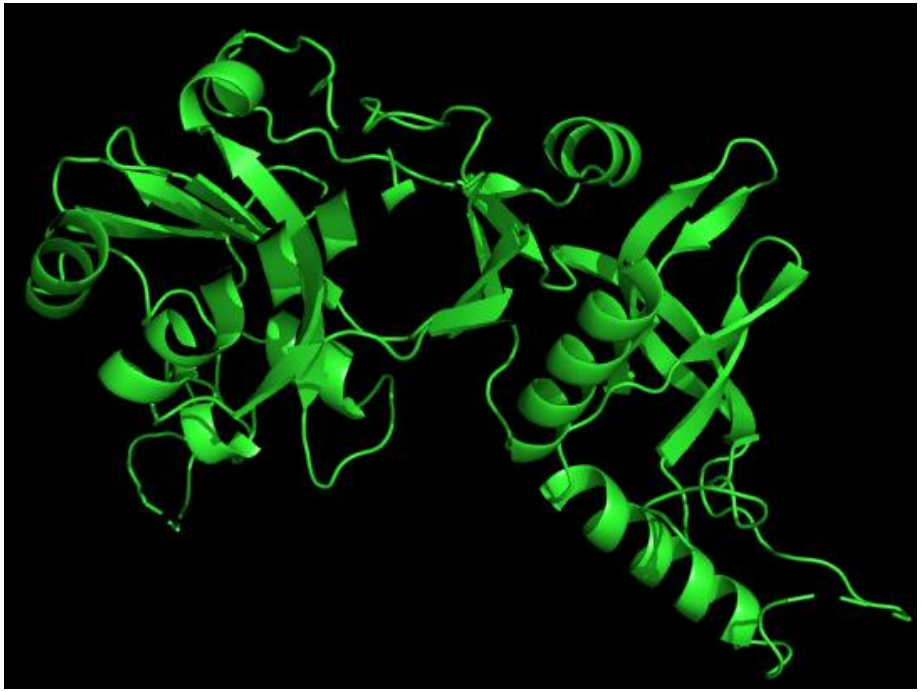
Fundamentals

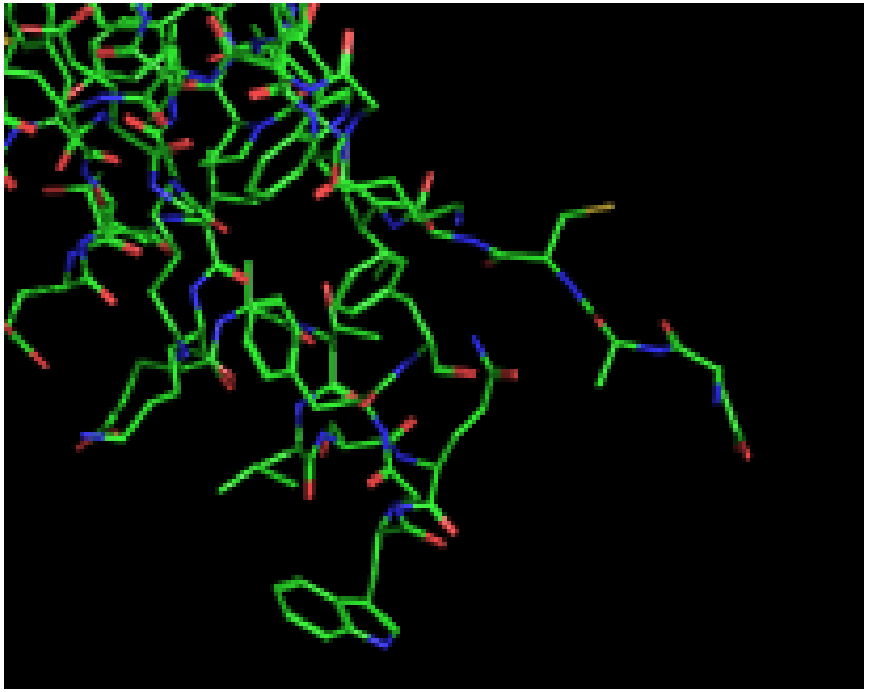


Fundamentals

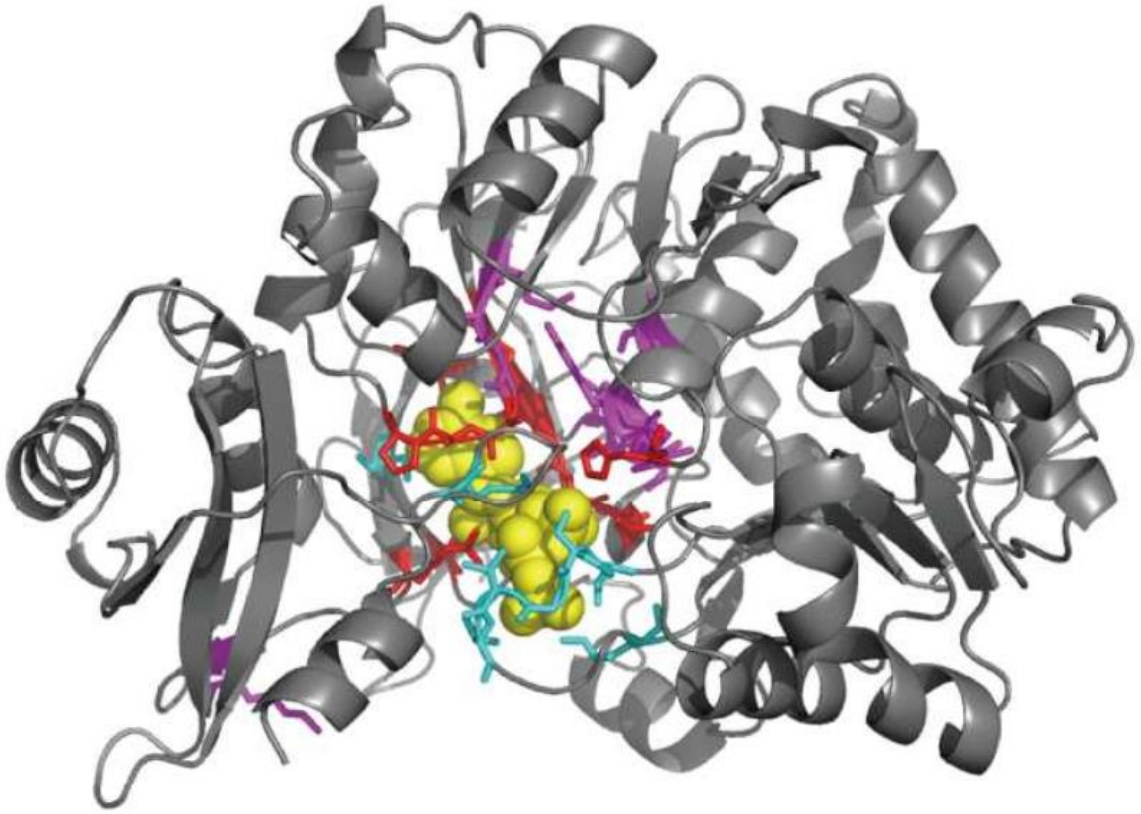
```
ATOM      1  N   VAL A   1     -10.053  -6.552  67.853  1.00 00.00           N
ATOM      2  CA  VAL A   1     -11.496  -6.584  67.619  1.00 00.00           C
ATOM      3  C   VAL A   1     -12.002  -7.947  68.142  1.00 00.00           C
ATOM      4  O   VAL A   1     -11.214  -8.539  68.885  1.00 00.00           O
ATOM      5  CB  VAL A   1     -11.850  -6.351  66.056  1.00 00.00           C
ATOM      6  CG1 VAL A   1     -10.934  -5.247  65.512  1.00 00.00           C
ATOM      7  CG2 VAL A   1     -11.726  -7.608  65.204  1.00 00.00           C
ATOM      8  N   ASN A   2     -13.187  -8.516  67.800  1.00 00.00           N
ATOM      9  CA  ASN A   2     -13.738  -9.702  68.488  1.00 00.00           C
ATOM     10  C   ASN A   2     -13.126 -11.020  68.031  1.00 00.00           C
ATOM     11  O   ASN A   2     -13.628 -11.790  67.211  1.00 00.00           O
ATOM     12  CB  ASN A   2     -15.274  -9.780  68.305  1.00 00.00           C
ATOM     13  CG  ASN A   2     -15.944 -10.692  69.337  1.00 00.00           C
ATOM     14  OD1 ASN A   2     -16.065 -10.347  70.515  1.00 00.00           O
ATOM     15  ND2 ASN A   2     -16.405 -11.875  68.953  1.00 00.00           N
```







Fundamentals



ATP in yellow.

TP(true positive) in red. **FP**(false positive) in magenta. **TN**(true negative) in gray. **FN**(false negative) in cyan.

Experiments

Statistical composition of the two data sets used in this study.

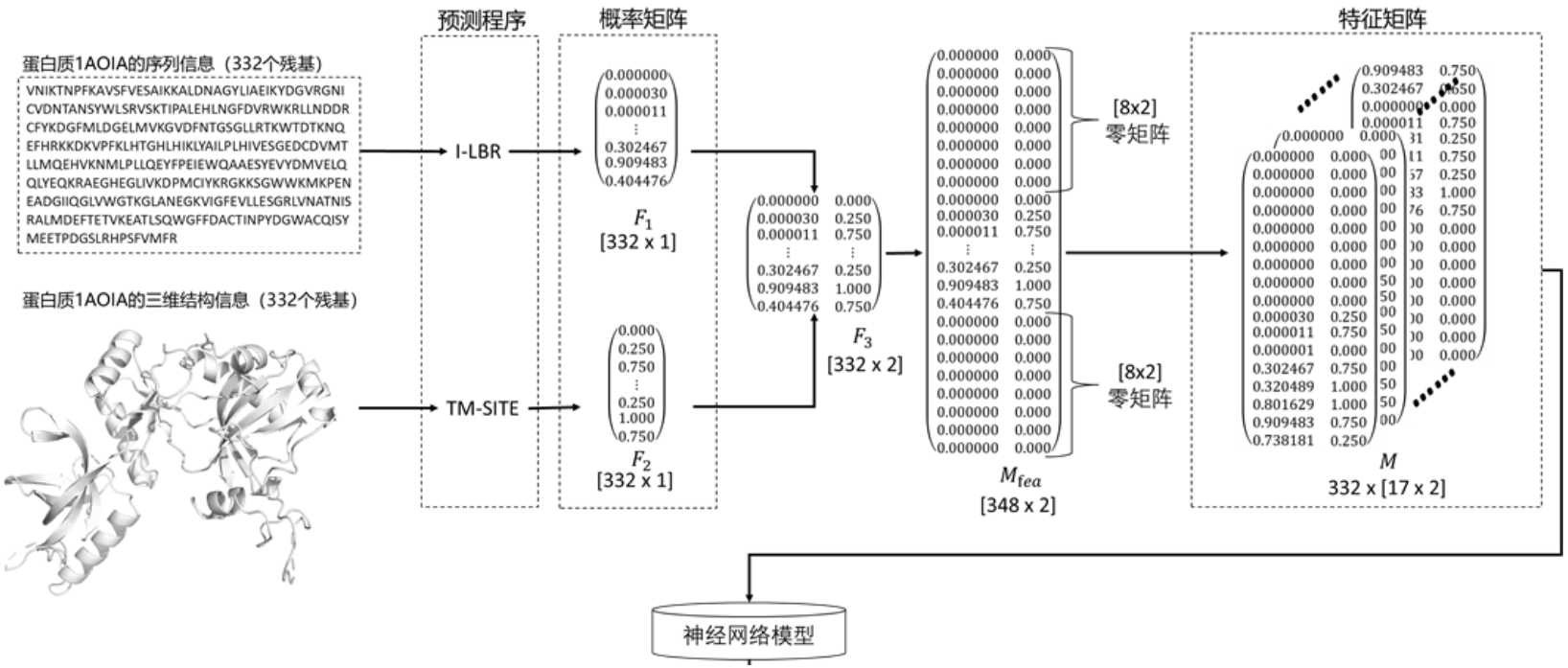
Data set	$N_{pro}{}^{a}$	$N_{pos}{}^{b}$	$N_{neg}{}^{c}$	$PNratio^{d}$
PART-388	388	5657	142,086	1: 25.12
PART-TEST	41	674	14,159	1: 21.01

[a] Number of proteins.

[b] Number of ATP-binding residues.

[c] Number of non-ATP-binding residues.

[d] $PNratio = N_{pos} : N_{neg}$.



Experiments

Statistical composition of the two data sets used in this study.

Data set	$N_{pro}{}^{a}$	$N_{pos}{}^{b}$	$N_{neg}{}^{c}$	$PNratio^{d}$
PART-388	388	5657	142,086	1: 25.12
PART-TEST	41	674	14,159	1: 21.01

[a] Number of proteins.

[b] Number of ATP-binding residues.

[c] Number of non-ATP-binding residues.

[d] $PNratio = N_{pos} : N_{neg}$.

features	feature number	Test dataset - 41					Training dataset - 388				
		Sen(%)	Spe(%)	Pre(%)	Acc(%)	Mcc	Sen(%)	Spe(%)	Pre(%)	Acc(%)	Mcc
PSFM	21	46.88	99.51	81.87	97.12	0.607					
ILBR	1	46.14	99.53	82.28	97.10	0.604					
TM-SITE	1	58.61	99.29	79.64	97.44	0.671					
PSFM+ILBR	22	56.38	99.10	74.80	97.16	0.635	79.80	99.65	89.94	99.89	0.841
PSFM+TM-SITE	22	62.46	99.13	77.39	97.47	0.682	81.74	99.55	87.91	98.87	0.842
TM-SITE+ILBR	2	57.27	99.49	84.28	97.57	**0.683**	55.03	99.43	79.31	97.73	0.650
PSFM+TM-SITE+ILBR	23	63.65	98.96	74.35	97.35	0.674	86.65	99.58	89.21	99.09	0.874

Experiments

Statistical composition of the two data sets used in this study.

Data set	$N_{pro}{}^{a}$	$N_{pos}{}^{b}$	$N_{neg}{}^{c}$	$PNratio^{d}$
PART-388	388	5657	142,086	1: 25.12
PART-TEST	41	674	14,159	1: 21.01

[a] Number of proteins.

[b] Number of ATP-binding residues.

[c] Number of non-ATP-binding residues.

[d] $PNratio = N_{pos} : N_{neg}$.

features	feature number	Test dataset - 41					Training dataset - 388				
		Sen(%)	Spe(%)	Pre(%)	Acc(%)	Mcc	Sen(%)	Spe(%)	Pre(%)	Acc(%)	Mcc
PSFM	21	46.88	99.51	81.87	97.12	0.607					
ILBR	1	46.14	99.53	82.28	97.10	0.604					
TM-SITE	1	58.61	99.29	79.64	97.44	0.671					
PSFM+ILBR	22	56.38	99.10	74.80	97.16	0.635	79.80	99.65	89.94	99.89	0.841
PSFM+TM-SITE	22	62.46	99.13	77.39	97.47	**0.682**	81.74	99.55	87.91	98.87	0.842
TM-SITE+ILBR	2	57.27	99.49	84.28	97.57	**0.683**	55.03	99.43	79.31	97.73	0.650
PSFM+TM-SITE+ILBR	23	63.65	98.96	74.35	97.35	0.674	86.65	99.58	89.21	99.09	0.874

Experiments

Statistical composition of the two data sets used in this study.

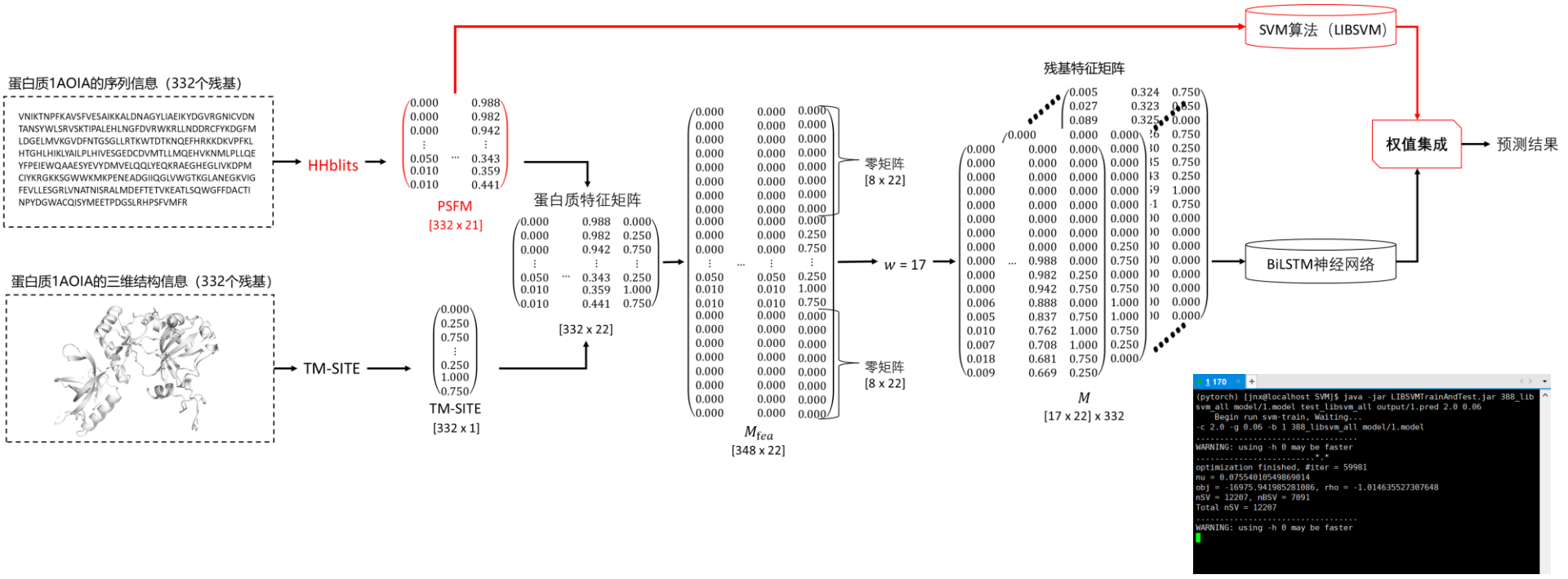
Data set	$N_{pro}{}^{a}$	$N_{pos}{}^{b}$	$N_{neg}{}^{c}$	$PNratio^{d}$
PART-388	388	5657	142,086	1: 25.12
PART-TEST	41	674	14,159	1: 21.01

[a] Number of proteins.

[b] Number of ATP-binding residues.

[c] Number of non-ATP-binding residues.

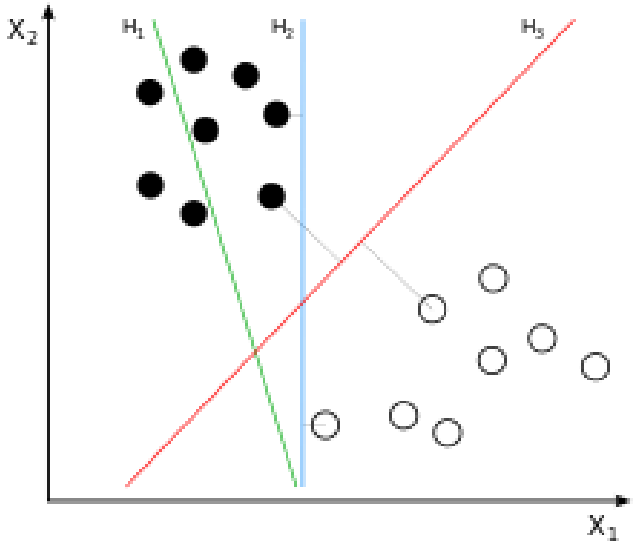
[d] $PNratio = N_{pos} : N_{neg}$.



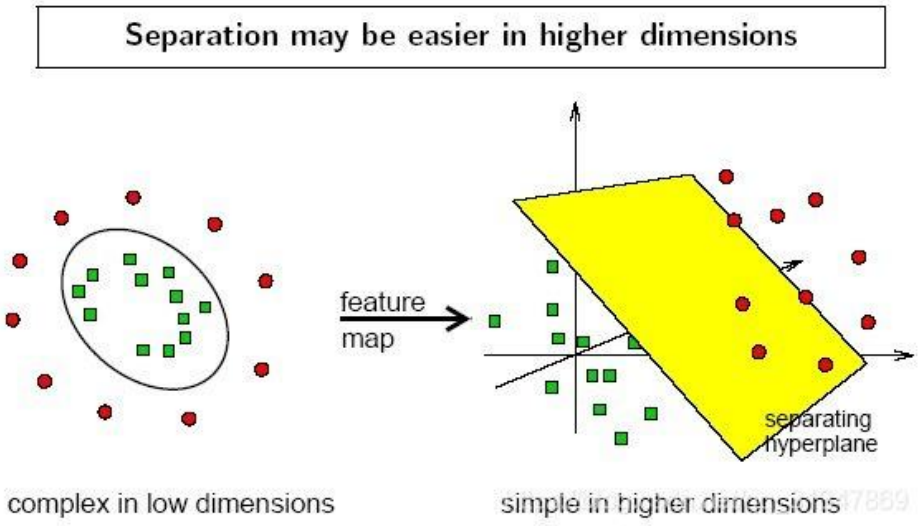

```
(pytorch) [jnx@localhost SVM]$ java -jar LIBSVMTrainAndTest.jar 388_lib
svm_all model/1.model test_libsvm_all output/1.pred 2.0 0.06
    Begin run svm-train, Waiting...
-c 2.0 -g 0.06 -b 1 388_libsvm_all model/1.model
.....................................
WARNING: using -h 0 may be faster
.....................*.*
optimization finished, #iter = 59981
nu = 0.07554010549869014
obj = -16975.941985281086, rho = -1.014635527307648
nSV = 12207, nBSV = 7091
Total nSV = 12207
.....................................
WARNING: using -h 0 may be faster
```

Network Structure

SVM（support vector machines）
支持向量机模型



线性可区分



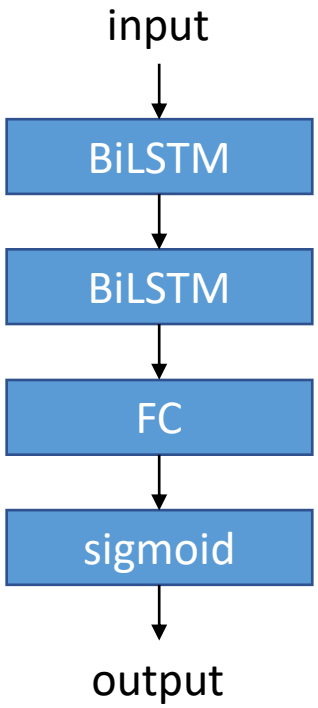
线性不可区分

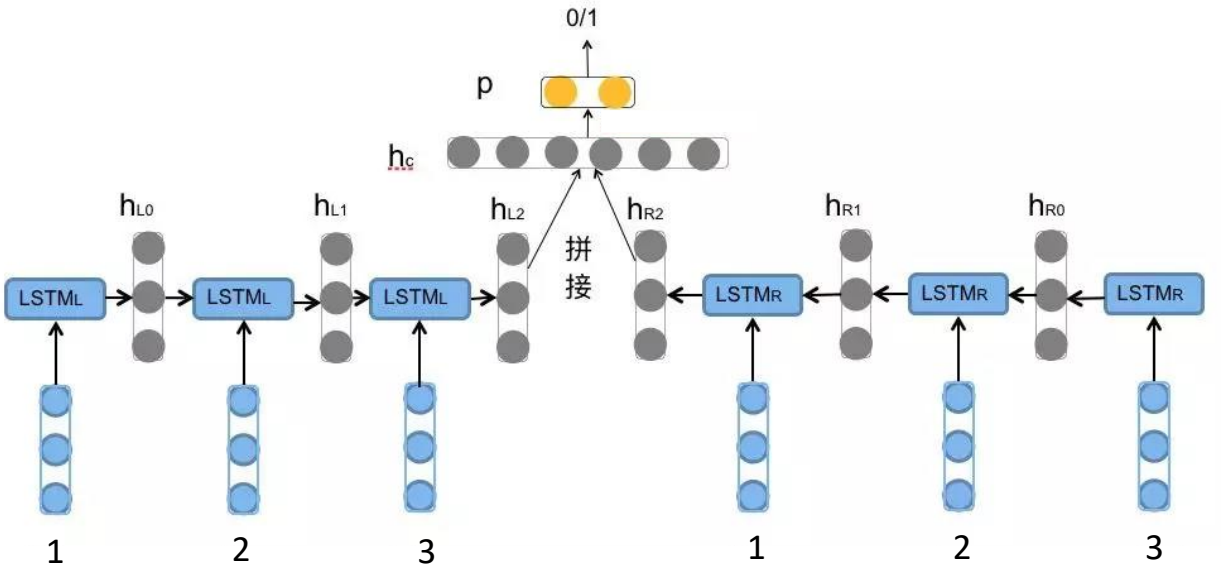
利用非线性映射将原始数据转化到高维空间中

Network Structure

BiLSTM（Bi-directional Long Short-Term Memory）

双向长短时神经网络

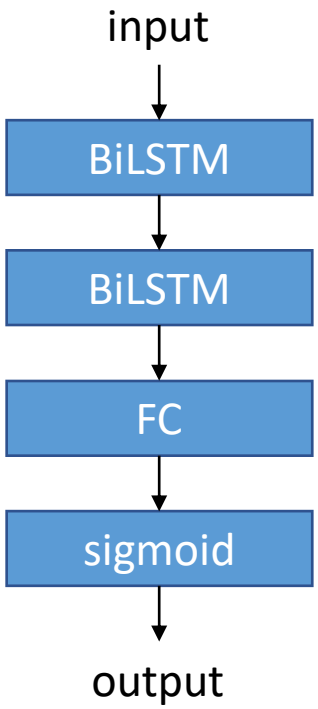


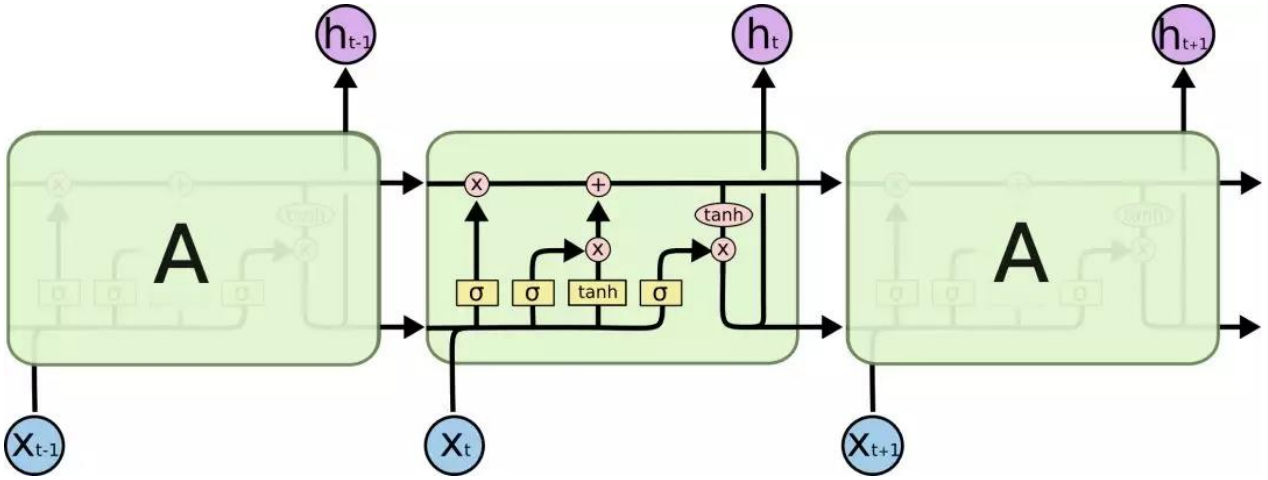


Network Structure

BiLSTM（Bi-directional Long Short-Term Memory）

双向长短时神经网络

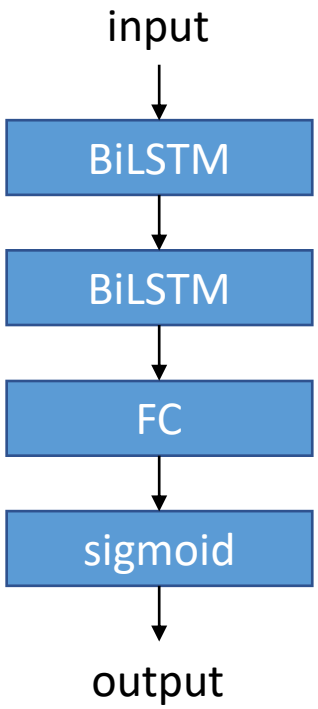


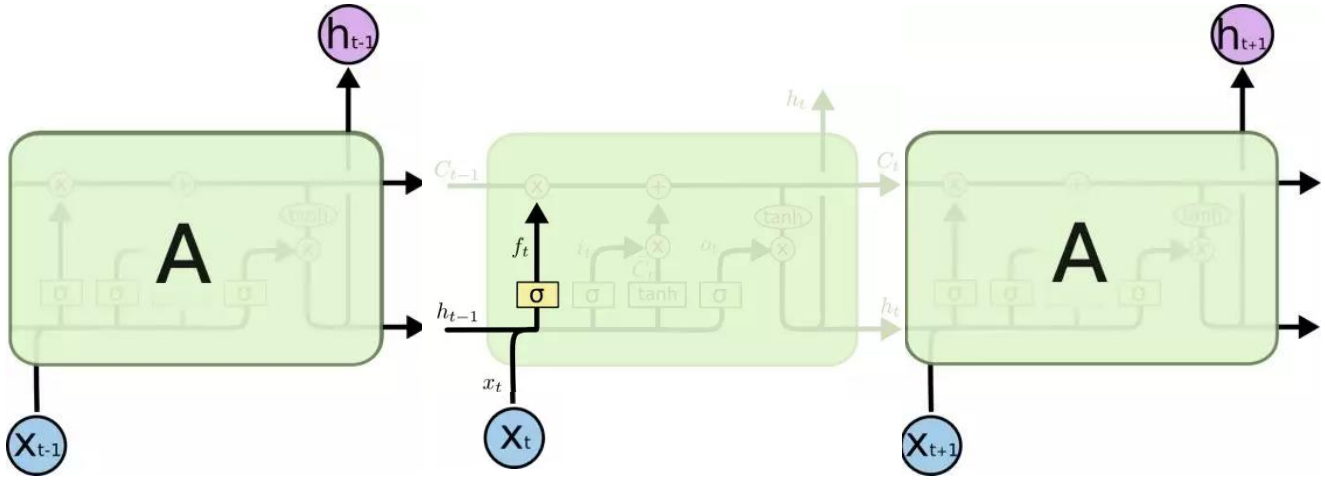


Network Structure

BiLSTM（Bi-directional Long Short-Term Memory）

双向长短时神经网络



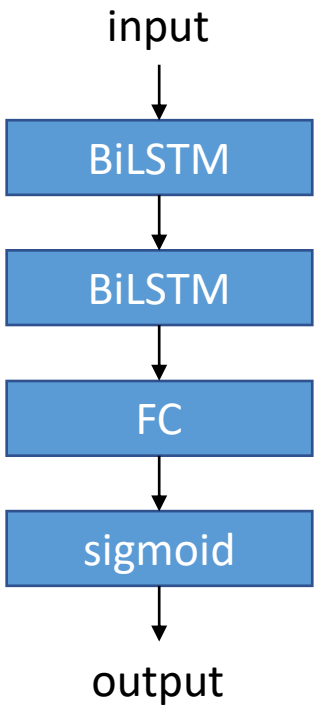


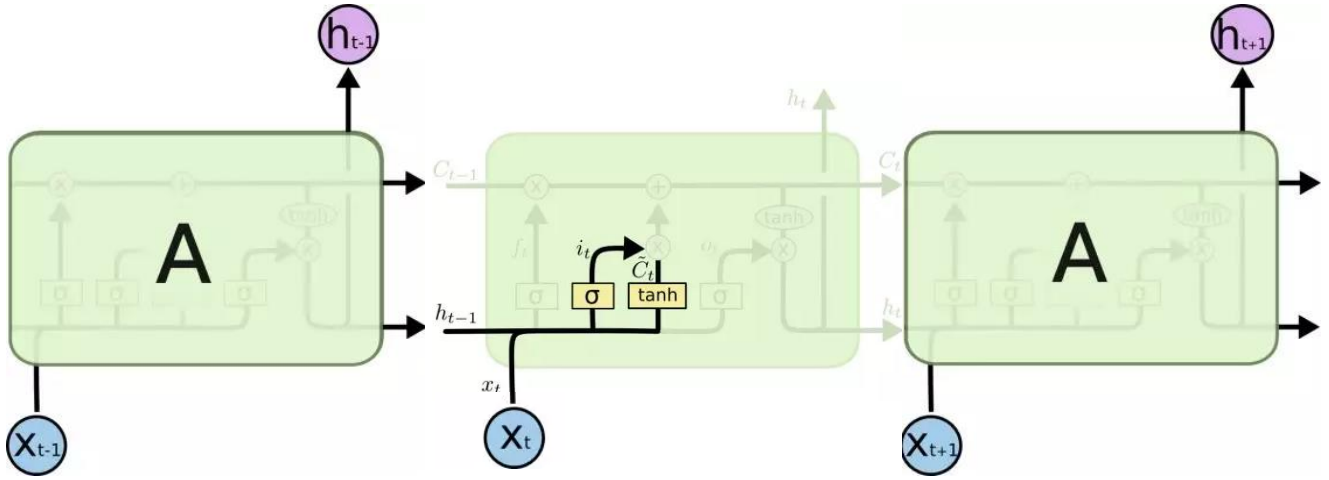
$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

Network Structure

BiLSTM（Bi-directional Long Short-Term Memory）

双向长短时神经网络





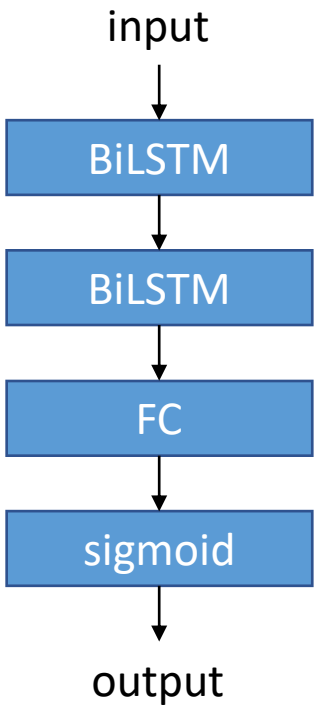
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

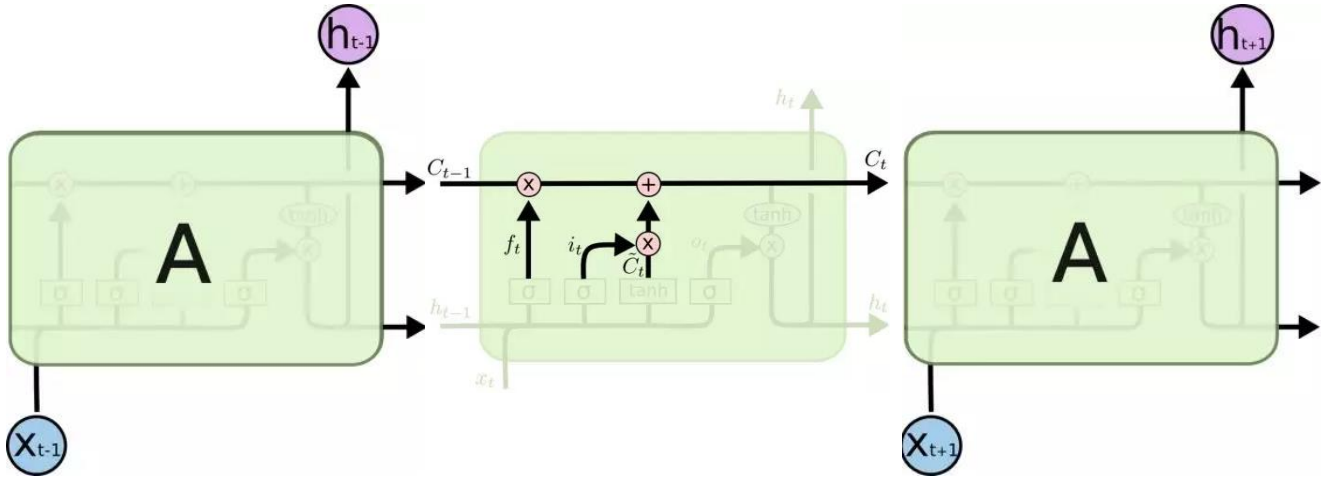
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

Network Structure

BiLSTM（Bi-directional Long Short-Term Memory）

双向长短时神经网络



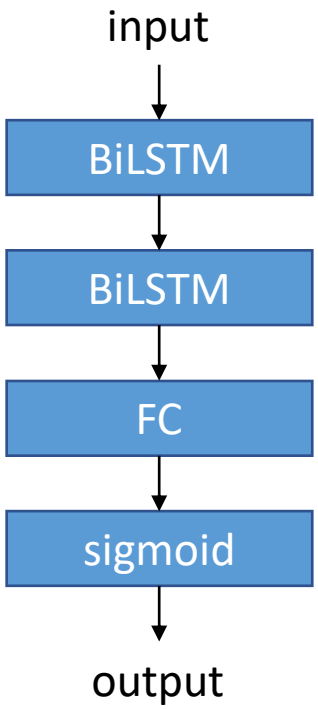


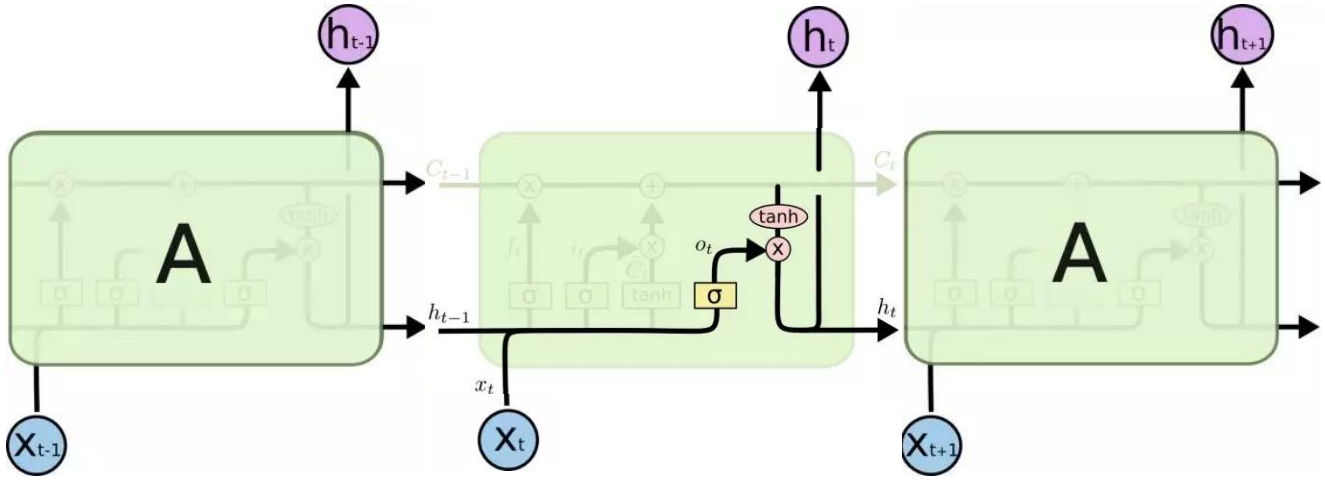
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Network Structure

BiLSTM（Bi-directional Long Short-Term Memory）

双向长短时神经网络





$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

Future work

1、模型集成

2、后处理实验