氨基酸序列
&
蛋白质结构

数据分析

汇报:薛智元

目录

S:载入训练数据

数据与赛制一栏提供下载：训练集文件*2

氨基酸序列文件（特征数据）：data_seq_train.txt

蛋白质二级结构文件（标签）：data_sec_train.txt

载入数据（代码）：

```
seq_df = pd.read_csv('./data_seq_train.txt',header = None)
sec_df = pd.read_csv('./data_sec_train.txt',header = None)
```

（两文件均无表头）

E：探索氨基酸/蛋白质数据

探索1

输入为共2W行的字符串

即2W条氨基酸链

输出亦为2W行的字符串

即2W个蛋白质链

1. 数据规模

```
print(seq_df.shape)
print(sec_df.shape)
```

```
(20000, 1)
(20000, 1)
```

2. df.head

```
print(seq_df.head)
```

```
<bound method NDFrame.head of
0      GPTGTGESKCPLMVKVLDAVRGSPAINVAVHVFRKAADDTWEPFAS...
1      MSHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDTHTAKYDPSLKP...
2      ADPGATLCLGHHAVPNGTLVKTITDDQIEVTNATELVQSSSTGKIC...
3         GSSGSSGTGEKPYKCNECGKAFRARSSLAIHQATHSGEKPSGPSSG
4      ENYARFITAASAARNPSPIRTMTDILSRGPKSMISLAGGLPNPNMF...
```

```
print(sec_df.head)
```

```
<bound method NDFrame.head of
0                  EEEEEEETTTTEE TT EEEEEEE TTSSEEEEEE...
1            SSTTTSGGGHHHH GGGGSSS S  EE TTTSEE TT   ...
2          EEEEEEEB  SS EEE  SS SSEEES EEE EE    SSEE...
3               S     EE TTS  EESSHHHHHHHHHH   S S TT
4         GGGS HHHHH                    EE       GGG ...
```

E：探索氨基酸/蛋白质数据

探索2

3.猜想：每行的输入与输出是否一一对应？True

4.字符的种类？

3. 20000条链中，seq与sec的对应关系？

```
flag = True
for i in range(20000):
    if len(str(seq_df.iat[i,0])) != len(str(sec_df.iat[i,0])):
        flag = False
        break
```

结果：True
结论：seq逐一对应sec，即每个氨基酸位均一一对应着一个蛋白质结构位

4. 统计氨基酸seq共多少种?蛋白质二级结构sec共多少种?

```
from collections import defaultdict
dic = defaultdict(int)

for i in range(20000):
    s = str(seq_df.iat[i,0])
    dic[s[i]] += 1
```

氨基酸seq共23种: 即26个英文字母中除去B，J，O
蛋白质二级结构sec共8种： [' ', 'B', 'E', 'G', 'I', 'H', 'S', 'T']

E：探索氨基酸/蛋白质数据

探索3

5.所有对应关系?

5. 每个seq可能对应多少种sec?

20种seq字符：

A C D E F
G H I K L
M N P Q R
S T V W Y

8种sec字符:

[' ', 'B', 'E', 'G', 'I', 'H', 'S', 'T']

剩余3种seq字符：

U

X

Z

不足8种sec字符：

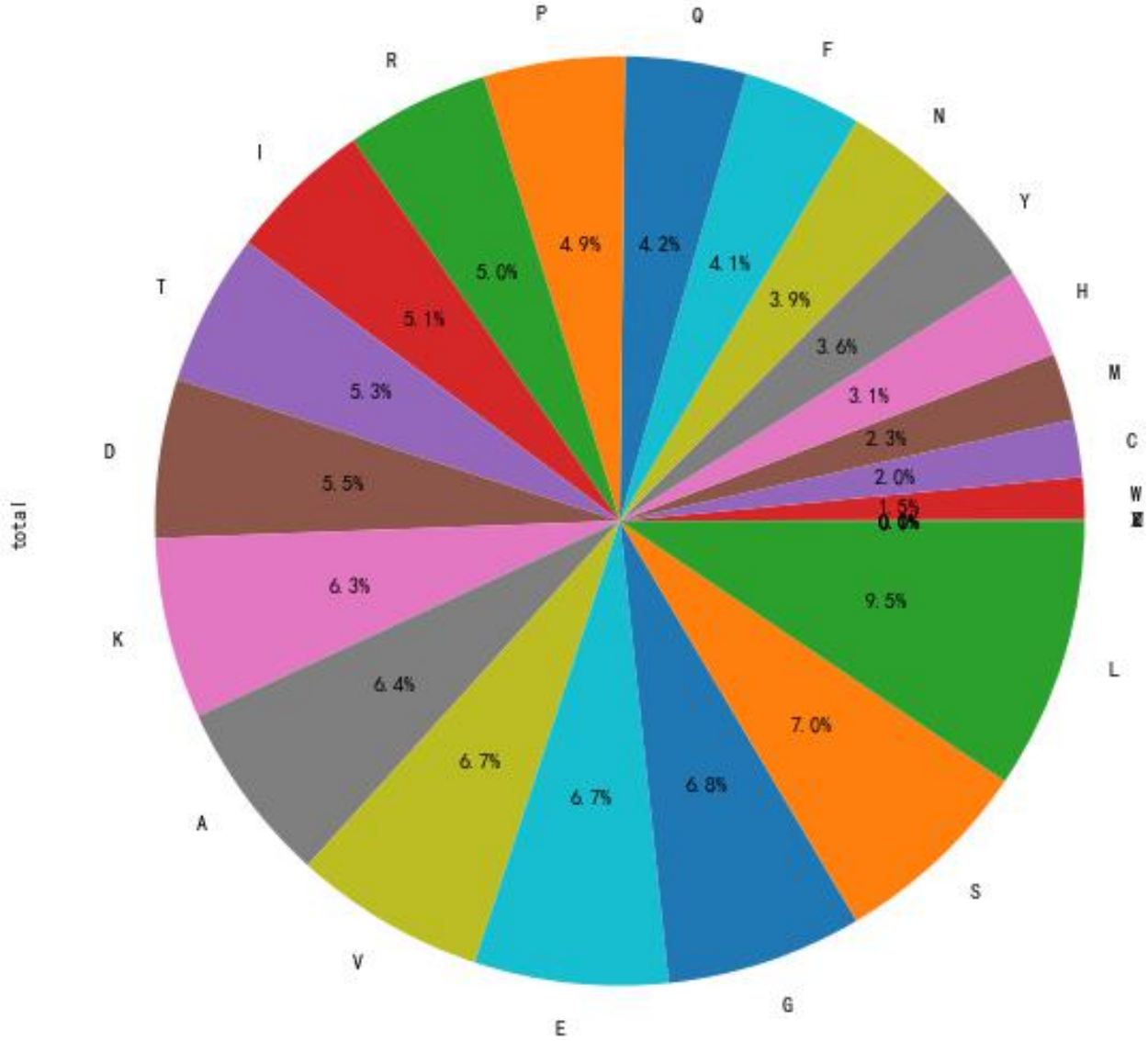
[' ', 'S']

[' ', 'B', 'E', 'G', 'H', 'S', 'T']

[' ']

E：探索氨基酸/蛋白质数据

6. 每种seq出现的频率?



探索4

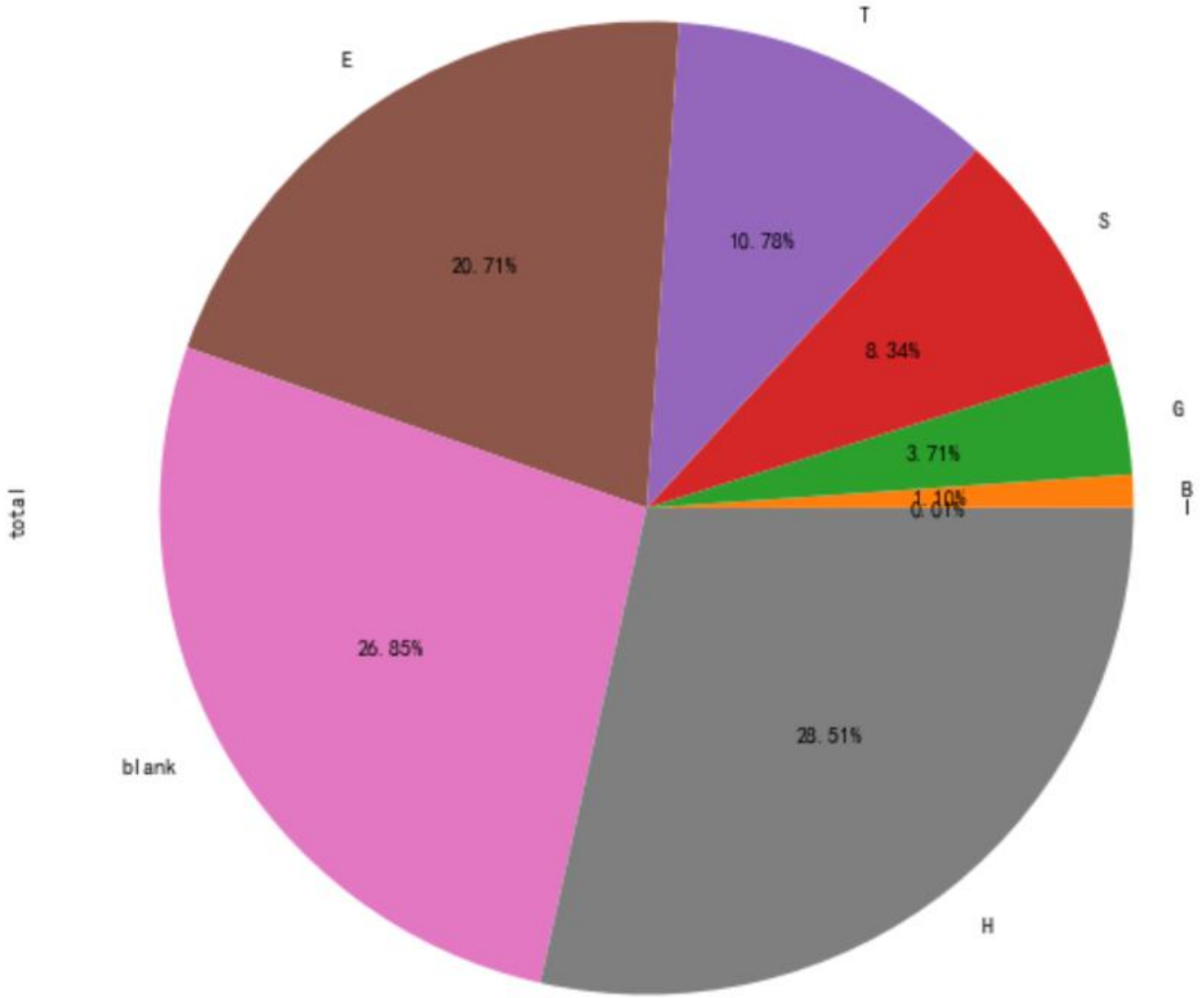
6.每种seq出现的频率

E：探索氨基酸/蛋白质数据

探索4

7.每种seq出现的频率

7. 每种sec出现的频率?



M&M 数据修正与建立模型

经查询领域内关于蛋白质二级结构预测的一般方法

可采取准确度较高的方法——PSSM：

引入氨基酸PSSM矩阵编码，

并利用LSTM模型训练每一个氨基酸位置，

并得到蛋白质二级结构结果，

PSSM谱码表如右图

1	S	2	2	−1	−4	−4	3	−1	−4	−4	−5	−5	2	−4	−6	−4	4	1	−6	−5	−4
2	T	−3	4	3	−4	−5	−1	−1	−4	−3	−5	−4	−1	2	−5	−5	3	5	−6	−5	−3
3	G	0	−1	−4	−2	−7	−5	−4	7	−6	−4	−7	−5	−7	−3	−5	0	−5	−7	−7	−5
4	S	0	−1	−1	−1	−5	0	0	−2	−2	−3	−4	−1	−4	−6	−2	5	3	−6	−3	−1
5	A	3	−5	−2	−1	−4	−3	−2	−1	−3	0	−2	−4	−2	−3	−3	1	2	−6	−2	4
6	T	1	−2	0	−1	−4	−1	−1	−2	−1	0	−3	−1	−1	−2	2	2	2	−5	−3	2
7	T	3	−6	1	−1	−4	−1	1	−2	−2	−1	0	−3	−1	−3	−3	1	2	−5	−1	1
8	T	−1	−2	0	−1	−1	0	1	−4	−1	1	−1	0	−1	−3	1	0	3	−2	−3	2
9	P	−3	−6	0	0	−6	−5	−5	−5	−4	−2	3	−4	−3	−4	7	0	−2	−6	−6	−2
10	I	−1	−3	−2	−3	−4	−1	−2	−4	0	3	2	−2	1	−1	−1	−1	1	−2	1	1
11	D	−2	−2	5	5	−7	2	1	−3	−2	−4	−6	−2	−6	−7	−3	−2	−1	−7	−3	−5
12	S⇨	0	−1	4	1	−5	0	−1	2	−1	−4	−5	−3	−3	1	−2	0	−3	−3	5	−3
13	L	−1	−2	2	−1	−4	1	0	−1	1	−2	1	−1	2	1	−3	0	0	1	2	−1
14	D	−4	−4	4	7	−8	0	0	−3	−3	−5	−7	−4	−5	−8	−7	−5	−5	−8	−5	5
15	D	1	−2	1	−1	−3	−2	−1	−2	−4	1	0	−3	1	0	−4	2	2	−5	1	2
16	A	1	−3	−4	−3	−6	4	3	−1	−4	−1	1	−3	1	0	−5	1	−2	−6	−1	2
17	V	9	9	8	10	4	9	9	10	5	7	8	9	8	2	8	9	9	7	10	7
18	T	−2	−4	−6	−8	−6	−4	−6	−7	−2	0	2	−6	0	5	−7	−2	0	−2	6	1
19	T	2	−7	−6	−6	2	−6	−6	5	−7	0	−4	−6	−1	−4	−7	1	1	−7	−4	1
20	P	−1	−2	3	1	−6	0	3	−5	−4	−2	−4	−1	−4	−7	4	0	2	−7	−3	2
21	V	−3	−8	−9	−9	−3	−7	−9	−5	−9	6	0	−8	0	−3	−8	−8	−4	−8	−7	5
22	Q	−1	−2	−1	−1	−4	3	0	1	−3	−5	−3	0	−3	−3	−6	3	3	−4	−1	5
23	I	−5	−9	−6	−9	−7	−8	−9	−8	−9	7	1	−8	0	1	−7	−8	−6	−8	−6	4
24	G	−6	−7	−6	−7	−9	−8	−7	8	−9	−10	−10	−8	−9	−10	−9	−7	−8	−9	−10	10
25	T	−4	−5	1	−2	−5	−4	−4	−5	−4	−6	−6	−4	−5	−6	−6	2	7	−6	−6	4
26	P	−4	−4	−1	−3	−6	−3	−3	−4	−3	−6	−6	−4	−6	−6	8	−2	−5	−6	−4	6

A：模型评价

模型评价方式——F1评分（2PR/(P+R)）

$$score = \frac{\sum_{i=0}^{n} F_1(A_i)}{n}$$

A_i为单个蛋白结构，

F_1函数为sklearn.metrics.f1_score，average='macro'


谢 谢 观 看