Term Project Final Report

หัวข้อโครงการ: การระบุตำแหน่งของ miRNA site บนสาย lncRNA ด้วย Deep Neural Network

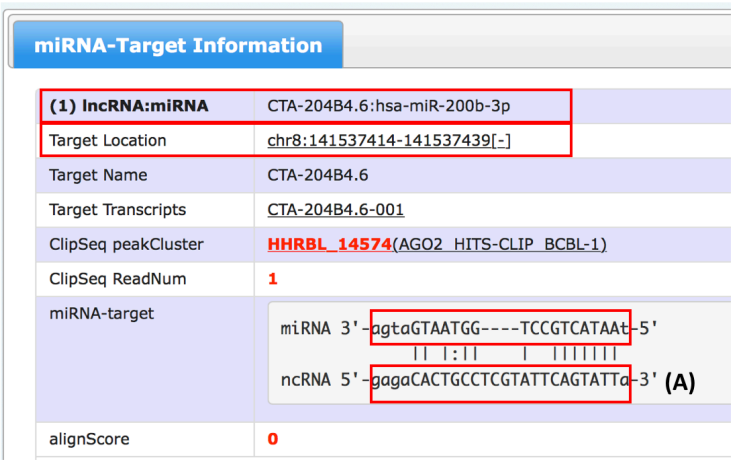
ผู้รับผิดชอบโครงการ: นายดนุภัทร คำนวนสินธุ์ รหัสนิสิต 5730196321

อาจารย์ที่ปรึกษาโครงการ: อ.ดร. ดวงดาว วิชาดากุล

โครงการนี้เป็นส่วนหนึ่งของวิชา 2110495 Bioinformatics ปีการศึกษาภาคปลาย 2560

ลักษณะข้อมูลที่ใช้

ใช้ข้อมูล miRNA site จากฐานข้อมูล starBase v2.0 และ lncipedia v3.1 เพื่อใช้ในการจัดเตรียมชุดข้อมูลสำหรับฝึกฝน และ ทดสอบ Deep Nueral Network



ภาพที่ 1 รูปแบบหน้าเว็ป starbase.sysu.edu.cn

ขั้นตอนการเตรียมข้อมูลมีดังนี้

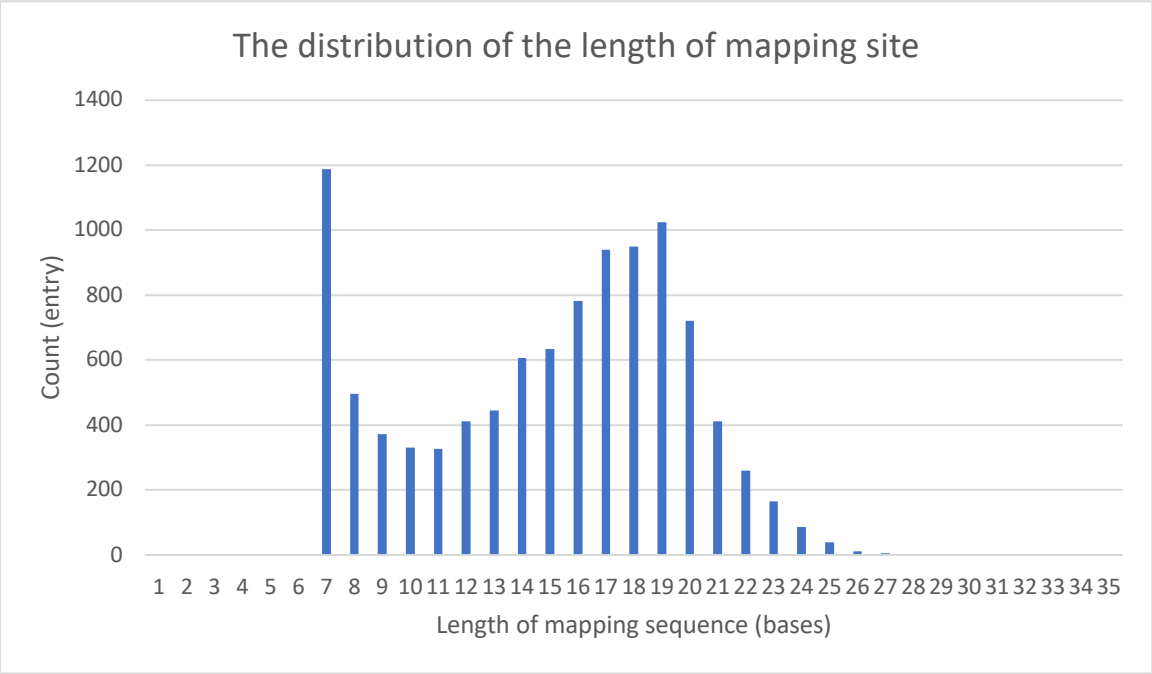
1. ทำการใช้ โปรแกรม web crawler ในการดาวน์โหลดข้อมูลจากเว็ป starbase.sysu.edu.cn ลงมาเก็บไว้ในรูปแบบ csv
2. ทำการใช้ข้อมูลในข้อที่ 1 ในการสร้างชุดข้อมูล lncRNA พร้อมกับ ตำแหน่งที่เกิด miRNA site และจัดเก็บลงในไฟล์รูปแบบ csv

ตัวอย่างของข้อมูลมีลักษณะดังนี้ โดยตัวอักษรสีแดงคือ ตำแหน่งที่เกิด miRNA site

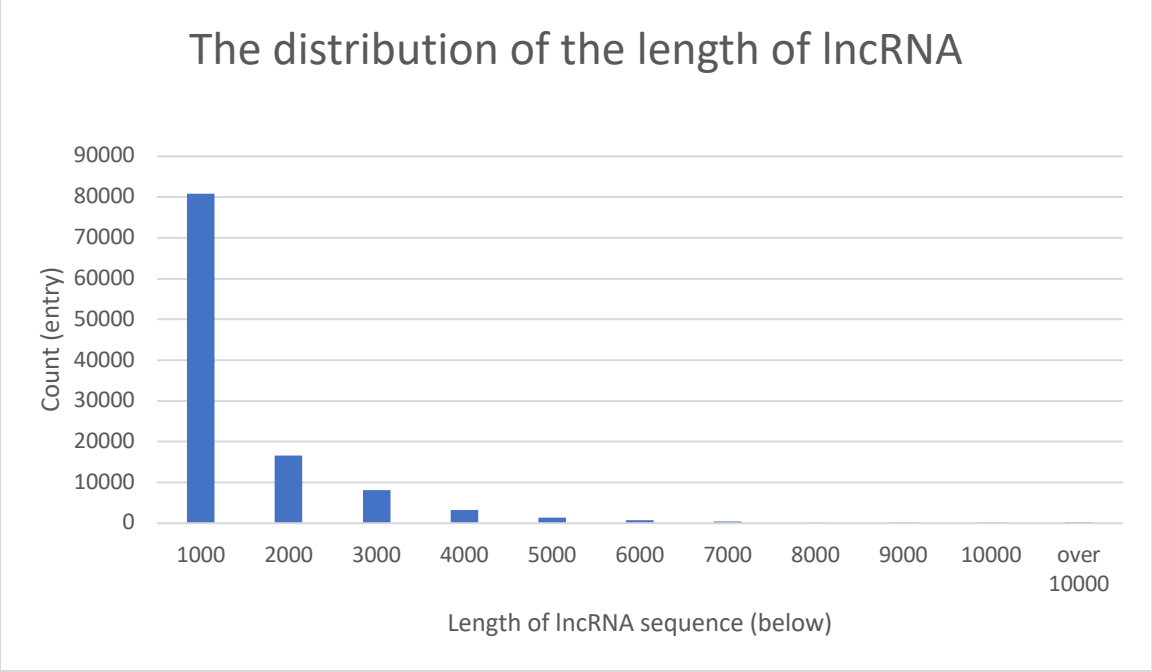
CTCTACACTACTGACCCAGGGCTTGGCAGTTTCTGTCCTAGAAATCAGAACACGC
AGCTCATGCATCACCGAGAAGGTGGTCAGCCTAAGAAGTGGTTCCCCACTCTTTGTTC
TAAGATATTCCTCTCTCAGA

รูปแบบการกระจายของข้อมูล

ความยาวของ miRNA site มีการกระจายตัวอยู่ในช่วง 7 ตัวอักษร ไปจนถึง 29 ตัวอักษร โดยพบเยอะที่สุดที่ 7 ตัวอักษร และ lncRNA ส่วนใหญ่มีความยาวไม่เกิน 10,000 เบส โดยรายละเอียดได้แสดงในภาพด้านล่างนี้



ภาพที่ 2 รูปแบบการกระจายของความยาวของ miRNA บน lncRNA



ภาพที่ 3 รูปแบบการกระจายของความยาวของ lncRNA

การจัดเตรียมชุดข้อมูล

เราสนใจเฉพาะ lncRNA ที่มีความยาวไม่เกิน 500 เบสเท่านั้น หาก lncRNA ใดมีความยาวไม่ถึง 500 เบส จะทำการ padding ด้วยตัวอักษร ‘P’ ต่อท้ายจนมีความยาวเท่ากับ 500 เบส โดยเมื่อทำการเลือกเอาเฉพาะ lncRNA ที่มี miRNA site ปรากฏอยู่ จากฐานข้อมูล starBase v2.0 พบว่ามีข้อมูล lncRNA ทั้งหมด 16,736 ข้อมูล โดยแบ่งออกเป็น

- ชุดข้อมูลสำหรับฝึกฝน จำนวน 60% ของข้อมูลทั้งหมด เท่ากับ 10,041 ข้อมูล
- ชุดข้อมูลสำหรับ Validate จำนวน 20% ของข้อมูลทั้งหมด เท่ากับ 3,348 ข้อมูล
- ชุดข้อมูลทดสอบ จำนวน 20% ของข้อมูลทั้งหมด เท่ากับ 3,348 ข้อมูล

โดยรูปแบบของข้อมูลมีดังนี้

- Raw_lncRNA : TGACCCAGGGCTTGGCAGTTTCTGTCCTAGAAATC
- X: [1,3,0,2,2,2,0,3,3,3,2,1,1,3,3,2,0,3,1,1,1,2,1,3,1,2,2,1,0,3,0,0,0,1,2]
- Y: [0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0]

โดยมีการแทนที่แต่ละตัวอักษรในส่วนของ X ดังนี้

- ตัวอักษร A จะถูกแทนที่ด้วยเลข 0
- ตัวอักษร T จะถูกแทนที่ด้วยเลข 1
- ตัวอักษร C จะถูกแทนที่ด้วยเลข 2
- ตัวอักษร G จะถูกแทนที่ด้วยเลข 3
- ตัวอักษร N จะถูกแทนที่ด้วยเลข 4
- ตัวอักษร P จะถูกแทนที่ด้วยเลข 5

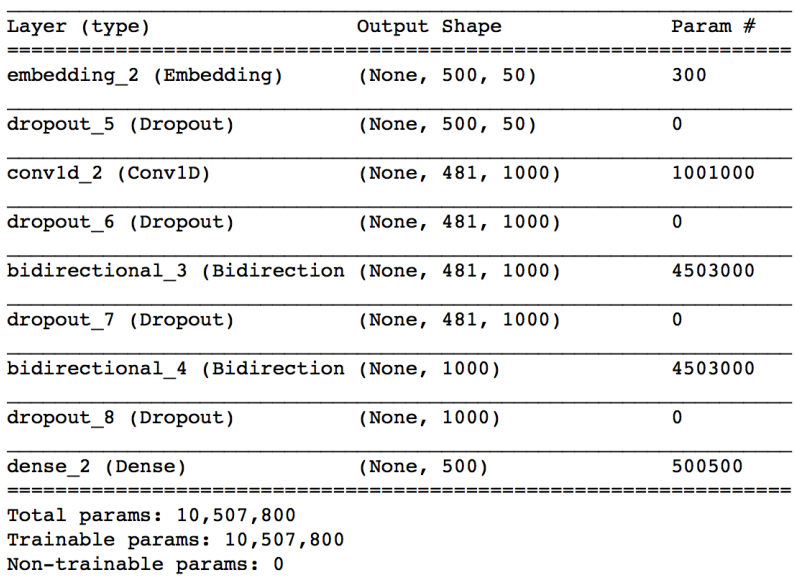
และในส่วนของ Y ดังนี้

- ส่วนที่ไม่เป็น miRNA site แทนที่ด้วยเลข 0
- ส่วนที่เป็น miRNA site แทนที่ด้วยเลข 1

หลักการ และ แนวคิดของ Deep Learning ของโครงการ

หลักการทำโครงการในครั้งนี้คือการ ทดลองใช้โมเดล deep neural network ในการเรียนรู้และจดจำลักษณะ หรือ รูปแบบ ของ miRNA site ที่เกิดขึ้นบนสาย lncRNA โดยตั้งสมมุติฐานไว้ว่า การเกิดขึ้นของ miRNA site บน lncRNA นั้นมีรูปแบบเฉพาะบางอย่างที่ซ่อนอยู่ โดยคาดหวังว่าโมเดล deep neural network จะสามารถค้นหาและเรียนรู้รูปแบบเฉพาะดังกล่าวที่ซ่อนอยู่ได้จากชุดข้อมูลฝึกฝนที่ได้จัดเตรียมขึ้น

โมเดล Deep Neural Network



Layer (type)	Output Shape	Param #
embedding_2 (Embedding)	(None, 500, 50)	300
dropout_5 (Dropout)	(None, 500, 50)	0
conv1d_2 (Conv1D)	(None, 481, 1000)	1001000
dropout_6 (Dropout)	(None, 481, 1000)	0
bidirectional_3 (Bidirection	(None, 481, 1000)	4503000
dropout_7 (Dropout)	(None, 481, 1000)	0
bidirectional_4 (Bidirection	(None, 1000)	4503000
dropout_8 (Dropout)	(None, 1000)	0
dense_2 (Dense)	(None, 500)	500500

Total params: 10,507,800
Trainable params: 10,507,800
Non-trainable params: 0

ภาพที่ 4 ภาพรวมของโมเดล Deep Nueral Network

1st Layer: Embedding Layer ขนาดข้อมูลนำเข้า 500 ตัวอักษร และ ขนาดของมิติเท่ากับ 50

เนื่องจากต้องการให้ model หาความสำพันธ์ระหว่างตัวอักษรแต่ละตัวด้วย (A, T, C, G, N, P) จึงใช้ Embedding Layer ในการทำ Character-based embedding ด้วยขนาดมิติเท่ากับ 50 ซึ่งมาจากผลการวิจัยของ Standford ในเอกสารประกอบ corpus GloVE ที่ระบุว่า ขนาดมิติตั้งแต่ 50 มิติขึ้นไป คือขนาดมิติขั้นต่ำที่เหมาะกับงานทางด้าน Embedding

2nd Layer: Dropout Layer

เพื่อลดการเกิด Overfitting กับชุดข้อมูลฝึกฝน การใส่ Dropout Layer จะช่วยให้โมเดลสามารถนำไปใช้กับชุดข้อมูลอื่นๆ ที่ไม่เคยเห็นมาก่อนได้ดียิ่งขึ้น

3rd Layer: 1-Dimension Convolutional Layer จำนวน 1,000 ฟิลเตอร์ ขนาด 20 ตัวอักษร

การใช้ Convolutional Layer จะทำให้โมเดลได้เรียนรู้ ความสัมพันธ์ระหว่าง 20 ตัวอักษร คล้ายการตัดในลักษณะ k-mers โดยขนาด 20 ตัวอักษรมาจากข้อมูลทางสถิติที่แสดงในกราฟที่ 1

4th Layer: Dropout Layer

5th Layer: Bi-directional-GRU จำนวน 1,000 หน่วย

หลังจากผ่าน Convolutional Layer มาแล้วจึงต้องการให้โมเดลจดจำรูปแบบของข้อมูลในลักษณะสองทาง กล่าวคือ ข้อมูลมีความเกี่ยวเนื่องกันทั้งข้อมูลก่อนหน้า และ ข้อมูลตามหลัง จึงเหมาะกับการนำ Recurrent Neural Network มาใช้งานในลักษณะของ Bi-directional

6th Layer: Dropout Layer

7th Layer: Bi-directional-GRU จำนวน 1,000 หน่วย

8th Layer: Dropout Layer

9th Layer: Output Layer จำนวน 1,000 หน่วย (จำนวนเท่ากับ input layer)

ชั้นสุดท้ายคือ Fully-connected Layer ที่มี Activation function คือ sigmoid function เนื่องจากต้องการให้ผลลัพธ์ออกมาในลักษณะ multi-class label คือสามารถระบุตำแหน่งของ miRNA site ได้โดยใน 1 สาย lncRNA สามารถมี miRNA site ได้มากกว่า 1 ตำแหน่ง

โดยเมื่อรวมทุก layer ของ deep neural network นี้เข้าด้วยกันจะมี parameter ให้สามารถเรียนรู้ได้ทั้งสิ้น 10,507,800 ค่า โดยทำการฝึกฝนจำนวน 15 epoch ใช้เวลาทั้งสิ้นประมาณ 4 ชั่วโมง

ผลลัพธ์

ผลการทดสอบในภาพรวมกับชุดข้อมูล Validate และ ชุดข้อมูลทดสอบที่จัดเตรียมไว้ได้ผลดังนี้

- ผลการทดสอบกับชุดข้อมูล Validate พบว่ามี loss เท่ากับ 0.0325 และ accuracy เท่ากับ 0.9880
- ผลการทดสอบกับชุดข้อมูลทดสอบพบว่ามี accuracy เท่ากับ 0.9888

เมื่อทำการพิจารณาเฉพาะส่วนที่เป็น miRNA site บนชุดข้อมูลทดสอบพบว่า

- สามารถระบุตำแหน่งของ miRNA site ได้ถูกต้องจำนวน 25,884 ตำแหน่ง จากทั้งหมด 44,615 ตำแหน่งโดยคิดเป็น 58%
- มีการระบุตำแหน่งผิดพลาดคือระบุว่าเป็น miRNA site แต่ในความเป็นจริงนั้นไม่ถูกต้อง จำนวน 8,651 ตำแหน่ง ทำให้ได้ค่าของ Precision เท่ากับ 0.7495
- มีตำแหน่งที่โมเดลไม่สามารถระบุได้ว่าเป็น miRNA site แต่ในความเป็นจริงนั้นเป็น miRNA site จำนวน 10,080 ตำแหน่ง ทำให้ได้ค่าของ Recall เท่ากับ 0.7197
- จากค่า Precision และ Recall ที่ได้ส่งผลให้ได้ F1-score เท่ากับ 0.7343

แนวทางในการพัฒนาต่อ

จากการทดลองใช้ Deep Neural Network ในการค้นหารูปแบบเฉพาะที่ซ่อนอยู่ที่ก่อให้เกิด miRNA site บน lncRNA นั้นพบว่ามีแนวโน้มที่สามารถทำได้ หากโมเดลได้รับชุดข้อมูลฝึกฝนที่มากขึ้นจะทำให้โมเดลมีแนวโน้มที่จะทำงานได้ดียิ่งขึ้น นอกจากนี้ยังสามารถประยุกต์โครงงานชิ้นนี้ไปใช้ในการสร้างการทำนาย miRNA บน lncRNA ที่มีความยาวมากกว่า 500 เบส ได้โดยใช้วิธีที่คล้ายกันนี้ โดยผู้พัฒนาได้เสนอแนวทางในการพัฒนาต่อเพื่อให้โมเดลสามารถทำนายตำแหน่งของ miRNA site บน lncRNA ขนาดมากกว่า 500 เบสไว้ดังด้านล่างนี้ ทั้งนี้การคาดการทั้งหมดอาจไม่เป็นจริง และ อาจทำไม่ได้หากไม่มีทรัพยากรในการประมวลผลที่เพียงพอ

แนวทางในการพัฒนาต่อ: การจัดเตรียมชุดข้อมูลสำหรับ lncRNA ที่มีความยาวมากกว่า 500

สามารถจัดเตรียมข้อมูลออกเป็น 4 รูปแบบดังนี้ซึ่งใช้ได้กับ lncRNA ที่มีความยาวเท่าใดก็ได้

- Backward
- Forward
- Cover
- Negative Sampling

ยกตัวอย่าง window_size มีค่าเท่ากับ 20 (ขนาดข้อมูลนำเข้าของ model)

Raw-sequence

AGAGGTGCCTCCCTTCCTTGAAACTTCTTCACCTTTGCTTCAAGAGCACCACAGAG
CAGATTCTTCAAGGCAGAGATATACAAGAATCCCTGTGGAAAACCTCATGAACTCAAGGTCAT
GCAGAGACTTAGGGAAGAACTGGGACTGGCTTATCAAGACTCTACACTACTGACCCAGGGCT
TGGCAGTTTCTGTCCTAGAAATCAGAACACGCAGCTCATGCATCACCGAGAAGGTGGTCAGC
CTAAGAAGTGGTTCCCCACTCTTTGTTCTAAGATATTCCTCTCTCAGA

Backward

CAGGGCTTGGCAGTTTCTGT
CCAGGGCTTGGCAGTTTCTG
CCCAGGGCTTGGCAGTTTCT
ACCCAGGGCTTGGCAGTTTC
GACCCAGGGCTTGGCAGTTT

TGACCCAGGGCTTGGCAGTT

CTGACCCAGGGCTTGGCAGT

ACTGACCCAGGGCTTGGCAG

TACTGACCCAGGGCTTGGCA

CTACTGACCCAGGGCTTGGC

ACTACTGACCCAGGGCTTGG

TACTACTGACCCAGGGCTTG

CTACTACTGACCCAGGGCTT

ACTACTACTGACCCAGGGCT

Forward

TTGGCAGTTTCTGTCCTAGA

TGGCAGTTTCTGTCCTAGAA

GGCAGTTTCTGTCCTAGAAA

GCAGTTTCTGTCCTAGAAAT

CAGTTTCTGTCCTAGAAATC

AGTTTCTGTCCTAGAAATCA

GTTTCTGTCCTAGAAATCAG

TTTCTGTCCTAGAAATCAGA

TTCTGTCCTAGAAATCAGAA

TCTGTCCTAGAAATCAGAAC

CTGTCCTAGAAATCAGAACA

TGTCCTAGAAATCAGAACAC

GTCCTAGAAATCAGAACACG

TCCTAGAAATCAGAACACGC

Cover

CTTGGCAGTTTCTGTCCTAG

GCTTGGCAGTTTCTGTCCTA

GGCTTGGCAGTTTCTGTCCT

GGGCTTGGCAGTTTCTGTCC

AGGGCTTGGCAGTTTCTGTC

Negative Sampling

CTACACTACTGACCCAGGGC

TCTACACTACTGACCCAGGG

CTCTACACTACTGACCCAGG

ACTCTACACTACTGACCCAG

GACTCTACACTACTGACCCA

.....

ภาคผนวก

Crawler Program

```
21
22  xl = pd.ExcelFile('doc/' + 'starBase_Human_Pan-Cancer_miRNA-LncRNA_Interactions2018-04-06_15-11.xlsx')
23  df = xl.parse("starBase_Human_Pan-Cancer_miRNA")
24
25  name = df['name']
26  geneName = df['geneName']
27
28  data_out = []
29
30  link_a = 'http://starbase.sysu.edu.cn/seedTargetInfo.php?type=lncRNA&database=hg19&name='
31  link_b = '&orgTable=mirLncRNAInteractionsAll'
32  autoid = 1
33
34  for i in range(len(name)):
35      try:
36          print('Crawl:',str(i+1)+'/'+str(len(name)))
37          query = name[i] + '&geneName='+ geneName[i] +'&autoId=' + str(autoid)
38          autoid += 1
39          r1 = requests.get(link_a+query+link_b)
40
41          html = r1.content.decode('utf8')
42          soup = BeautifulSoup(html,"lxml")
43          text = soup.get_text()
44
45          lncRNA_pos = text.index('lncRNA:miRNA')
46          target_pos = text.index('Target Location')
47          end_target_pos = text.index('Target Name')
48          miRNA_pos = text.index('miRNA-targetmiRNA')
49          end_miRNA_pos = text.index('alignScore')
50
51          lncRNA = text[lncRNA_pos+12:target_pos].replace('\n','')
52          target_location = text[target_pos+15:end_target_pos].replace('\n','')
53
54          miRNA_target = text[miRNA_pos+17:end_miRNA_pos].strip().split('\n')
55          miRNA = miRNA_target[0]
56          ncRNA = miRNA_target[2]
57
58          data_out.append((lncRNA, target_location, miRNA, ncRNA))
59      except:
60          print('Error:',str(i+1)+'/'+str(len(name)))
61          save(data_out,'starbase_crawl-bak.pkl')
62          pass
63
64  print('Collected:',len(data_out))
65  save(data_out,'starbase_crawl.pkl')
```

Training Logs

```
Train on 10041 samples, validate on 3348 samples
Epoch 1/30
10016/10041 [============================>.] - ETA: 2s - loss: 0.1100 -
acc: 0.9754Epoch 00000: val_loss improved from inf to 0.09086, saving m
odel to stat/w/weights-jet.h5
10041/10041 [==============================] - 940s - loss: 0.1100 - ac
c: 0.9754 - val_loss: 0.0909 - val_acc: 0.9793
Epoch 2/30
10016/10041 [============================>.] - ETA: 2s - loss: 0.0822 -
acc: 0.9787Epoch 00001: val_loss improved from 0.09086 to 0.06970, savi
ng model to stat/w/weights-jet.h5
10041/10041 [==============================] - 944s - loss: 0.0822 - ac
c: 0.9787 - val_loss: 0.0697 - val_acc: 0.9815
Epoch 3/30
10016/10041 [============================>.] - ETA: 2s - loss: 0.0711 -
acc: 0.9803Epoch 00002: val_loss improved from 0.06970 to 0.06335, savi
ng model to stat/w/weights-jet.h5
10041/10041 [==============================] - 944s - loss: 0.0711 - ac
c: 0.9803 - val_loss: 0.0633 - val_acc: 0.9822
Epoch 4/30
10016/10041 [============================>.] - ETA: 2s - loss: 0.0588 -
acc: 0.9824Epoch 00003: val_loss improved from 0.06335 to 0.05366, savi
ng model to stat/w/weights-jet.h5
10041/10041 [==============================] - 936s - loss: 0.0588 - ac
c: 0.9824 - val_loss: 0.0537 - val_acc: 0.9838
Epoch 5/30
10016/10041 [============================>.] - ETA: 2s - loss: 0.0568 -
acc: 0.9812Epoch 00004: val_loss improved from 0.05366 to 0.04898, savi
ng model to stat/w/weights-jet.h5
10041/10041 [==============================] - 939s - loss: 0.0568 - ac
c: 0.9812 - val_loss: 0.0490 - val_acc: 0.9848
Epoch 6/30
10016/10041 [============================>.] - ETA: 2s - loss: 0.0462 -
acc: 0.9841Epoch 00005: val_loss improved from 0.04898 to 0.04473, savi
ng model to stat/w/weights-jet.h5
10041/10041 [==============================] - 947s - loss: 0.0462 - ac
c: 0.9842 - val_loss: 0.0447 - val_acc: 0.9858
Epoch 7/30
10016/10041 [============================>.] - ETA: 2s - loss: 0.0401 -
acc: 0.9858Epoch 00006: val_loss improved from 0.04473 to 0.04323, savi
ng model to stat/w/weights-jet.h5
10041/10041 [==============================] - 951s - loss: 0.0401 - ac
c: 0.9858 - val_loss: 0.0432 - val_acc: 0.9862
Epoch 8/30
10016/10041 [============================>.] - ETA: 2s - loss: 0.0371 -
acc: 0.9867Epoch 00007: val_loss improved from 0.04323 to 0.04002, savi
ng model to stat/w/weights-jet.h5
10041/10041 [==============================] - 945s - loss: 0.0371 - ac
c: 0.9867 - val_loss: 0.0400 - val_acc: 0.9874
Epoch 9/30
10016/10041 [============================>.] - ETA: 2s - loss: 0.0353 -
acc: 0.9872Epoch 00008: val_loss did not improve
10041/10041 [==============================] - 946s - loss: 0.0353 - ac
c: 0.9872 - val_loss: 0.0471 - val_acc: 0.9853
```

```
Epoch 10/30
10016/10041 [============================>.] - ETA: 2s - loss: 0.0420 -
acc: 0.9853Epoch 00009: val_loss did not improve
10041/10041 [==============================] - 942s - loss: 0.0420 - ac
c: 0.9853 - val_loss: 0.0519 - val_acc: 0.9836
Epoch 11/30
10016/10041 [============================>.] - ETA: 2s - loss: 0.0374 -
acc: 0.9867Epoch 00010: val_loss did not improve
10041/10041 [==============================] - 946s - loss: 0.0374 - ac
c: 0.9867 - val_loss: 0.0422 - val_acc: 0.9872
Epoch 12/30
10016/10041 [============================>.] - ETA: 2s - loss: 0.0330 -
acc: 0.9879Epoch 00011: val_loss did not improve
10041/10041 [==============================] - 949s - loss: 0.0330 - ac
c: 0.9879 - val_loss: 0.0409 - val_acc: 0.9877
Epoch 13/30
10016/10041 [============================>.] - ETA: 2s - loss: 0.0335 -
acc: 0.9880Epoch 00012: val_loss did not improve
10041/10041 [==============================] - 950s - loss: 0.0335 - ac
c: 0.9880 - val_loss: 0.0425 - val_acc: 0.9874
Epoch 14/30
10016/10041 [============================>.] - ETA: 2s - loss: 0.0328 -
acc: 0.9881Epoch 00013: val_loss did not improve
10041/10041 [==============================] - 951s - loss: 0.0328 - ac
c: 0.9881 - val_loss: 0.0427 - val_acc: 0.9877
Epoch 15/30
10016/10041 [============================>.] - ETA: 2s - loss: 0.0325 -
acc: 0.9881Epoch 00014: val_loss did not improve
10041/10041 [==============================] - 949s - loss: 0.0325 - ac
c: 0.9881 - val_loss: 0.0409 - val_acc: 0.9880
```

Project Repository

Github: https://github.com/jrkns/bioclass/tree/master/term_project