DNA-蛋白質的相連預測

研究背景

基因體(Genome)內所包含的遺傳訊息，對於生物的演化以及遺傳疾病的研究有著相當重要的影響，而染色質的結構會影響基因表現、蛋白質表現、生物途徑等等。在基因的某些區間中，稱作 open region，可以和轉錄因子(transcription factors，一種蛋白質)、或是 RNA 聚合酶相連產生表現，相反的，有些區間則是緊密封閉，和基因表現無關。現在有許多高通量，全基因組的分析方法如 DNase-seq、FAIRE-seq 和 ATAC-seq，用來尋找基因中 open region，但這些方法都有相同的問題，就是費用昂貴且耗時間。因此，近年來有了用演算法或是深度學習來找出 open region 的方法出現。

研究目的

1. 學習，熟悉 CNN 的使用
2. 訓練作研究的流程與方法

研究過程

1. 資料取得 ： 從 ENCODE project 官網，選定叫 CTCF 的蛋白質，下載和 CTCF 相連的 DNA 序列(所有的序列長度都相等)，作為資料中的 postive sample，而所有的 negative sample 則是把 postive 的 DNA 序列順序打亂，如原本是 ACTCA，打亂後變成 CCTAA，所有的序列長度都是 101。
2. 資料前處理 ： 為了將 DNA 序列轉換成 CNN 能處理的形式，使用 one-hot encoding 的方法，將序列中的 ACTG 鹼基，按照 DNA 序列的順序轉換，如果原本是 ACTGTC，則會變成如下

[1,0,0,0
0,1,0,0
0,0,1,0
0,0,0,1
0,0,1,0
0,1,0,0]

3. Model 架構

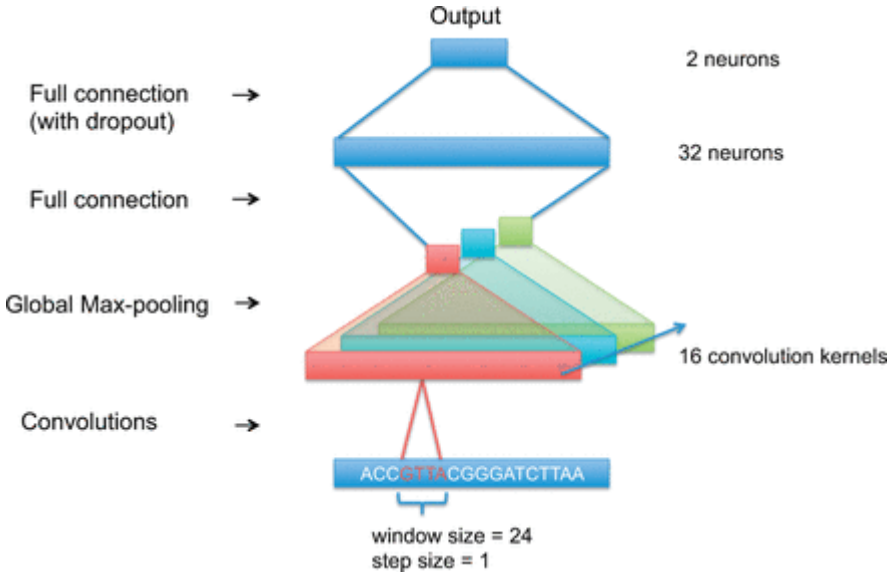


示意圖 出自[1]

如上圖所示，轉換成矩陣形式的資料，經過 CNN 處理，再將輸出的結果放入 full connection 中，最後 output 的結果會是 0 或 1，0 代表 model 預測此 DNA 序列是 negative，反之 1 是 positive 。

4. 實驗結果

最初的結果 test data 正確率一直停在 50%，代表 model 的訓練並不成功，因為用猜的正確率也會接近 50%，後來在改變 CNN 的 filter size 後有了較大的突破，原本的 filter size 用比較小，像是[2,2] 或 [3,3] ，但發現使用[20,4]的 filter size 才能訓練出好的結果，最終 test data 的正確率可達到 88.5%，推測是 DNA 序列中和蛋白質相連的區域有一定的長度，太小的 filter 會找不到這些區域。

參數設置

batch size = 350

training step = 12000

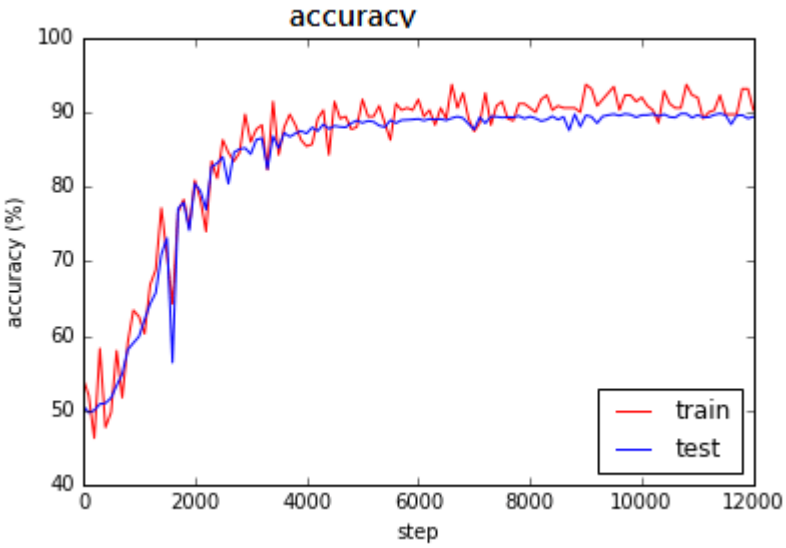
learning rate = 0.25 (starting) with exponential decay after 5000 steps, decay rate = 0.5

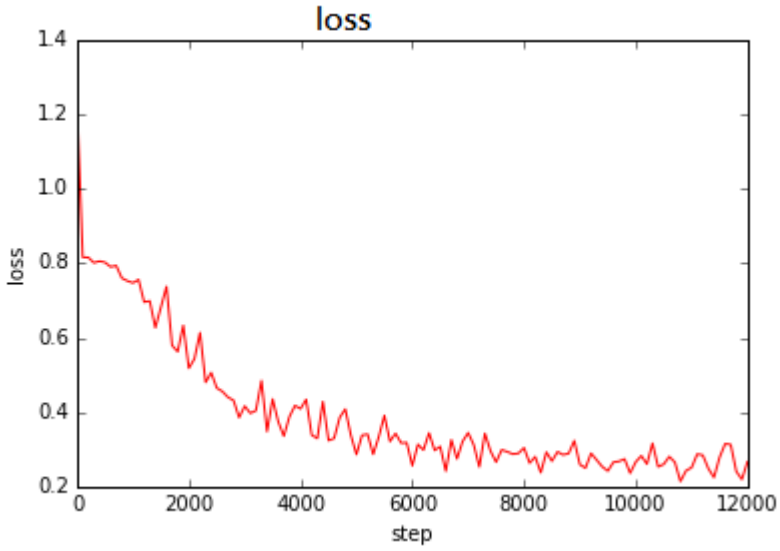
optimizer: GradientDescentOptimize

dropout rate = 0.5

convolution filter size: 第一層[20,4] 第二層[8,4]

正確率和 loss 隨 step 的變化圖





5.Reference

[1] Convolutional neural network architectures for predicting DNA-protein binding https://academic.oup.com/bioinformatics/article/32/12/i121/2240609#84802981

6.相關論文

Bioinformatics, 33, 2017, i92–i101
doi: 10.1093/bioinformatics/btx234
ISMB/ECCB 2017

OXFORD

Chromatin accessibility prediction via convolutional long short-term memory networks with *k*-mer embedding

Xu Min[1,2], Wanwen Zeng[1,3], Ning Chen[1,2], Ting Chen[1,2,4,*] and Rui Jiang[1,3,*]

[1]MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST, [2]Department of Computer Science and Technology, State Key Lab of Intelligent Technology and Systems, [3]Department of Automation, Tsinghua University, Beijing 100084, China and [4]Program in Computational Biology and Bioinformatics, University of Southern California, CA 90089, USA

*To whom correspondence should be addressed.

簡介：這篇論文發表在 Bioinformatics 2017,7 是很新的論文，作者不再把 DNA 序列用 one-hot encoding 的方法轉換成矩陣，而是把 “A” “C” “T” “G” 分別看成文字，而由 ACTG 組成的序列則當作句子，用 Glove 套件訓練出 embedding，再放到 CNN 中，而為了可以處理不同序列長度，CNN 後的結果會放到 LSTM 中，最後才會輸出預測結果。

$$\mathbf{h} = g(\mathbf{x}) = g_{\text{lstm}}(g_{\text{conv}}(g_{\text{embed}}(\mathbf{x}))).$$