HW2 - Estimate epitope residues in protein sequence using CNN network.

The homework is to be submitted in groups up to 4 students .

Input:

Protein Sequence - $a_1, a_2, \ldots, a_N$. a_i - Amino acid #i in chain

Output:

$p(a_1), p(a_2), \ldots, p(a_N)$, $p(a_k)$—probability of the amino acid k to be in the epitope

Preprocessing:

- Process for each residue
  - Calculate :
    - computed volume
    - hydrophobicity
    - polarity
    - relative surface accessibility (RSA)
    - secondary structure (SS)
    - Type

CNN :

- ✦ Extract window of N>9 amino acids
- ✦ Build a CNN network :
  - ✦ Input: Nx6 matrix of amino acid properties
  - ✦ Output: $p(a_{N/2})$ - the probability of the center amino acid to be in epitopes
- ✦ Make at least three hidden layers

Training, Testing, Analysis Discussion:

- Initial train CNN. Choose the datasets such that the amounts of positive (epitopes) and negative entries are approximately equal :
  - Training DataSet of 300 **windows,** Testing DataSet of 300 windows:
    1. Show Train Loss, Test Loss, Train Accuracy, Test Acuuracy. Does the CNN overfit
  - Training DataSet of 30000 windows**,** Testing DataSet of 3000 windows. :
    2. Show Train Loss, Test Loss, Train Accuracy, Test Acuuracy. Does the CNN overfit
- **Sigmoid vs Softmax** : Try two different last layer configurations: Sigmoid(logistic), i.e. and Softmax, i.e. 2-D output. Training DataSet of 30000 windows**,** Testing DataSet of 3000 windows:
    3. Show on the same graph Train Loss for Sigmoid and Softmax Configuration.
    4. Show on the same graph Test Loss for Sigmoid and Softmax Configuration.
    5. Show on the same graph Test Accuracy for Sigmoid and Softmax Configuration.
- **CNN configurations**
  - Choose the configuration with better results from the previous paragraph
    - Add one convolution layer and compare to the original configuration
      6. Show on the same graph Train Loss for two configurations
      7. Show on the same graph Train Loss vs actual time for two configurations
      8. Show on the same graph Test Loss for two configurations
      9. Show on the same graph Test Accuracy for r two configurations
  - Choose the configuration with better results from the previous paragraph
    - Add one fully connected layer and compare to the original configuration
      10. Show on the same graph Train Loss for two configurations
      11. Show on the same graph Train Loss vs actual time for two configurations
      12. Show on the same graph Test Loss for two configurations
      13. Show on the same graph Test Accuracy for r two configurations

Submission:

- To Moodle in single zip file (one per group)
  - PDF document
    - List of students ID's and mails
    - Brief Description of all Networks Used
    - Hardware Used
    - Graphs 1-13 in a readable form (ticks, legends, captions)
  - Preproccesing Code
  - Torch Code

Database:

- Use linear epitope database for train and test http://www.cbs.dtu.dk/services/BepiPred/download.php

References:

PDB format - https://www.rcsb.org/pdb/static.do?p=file_formats/pdb/index.html
Chimera Tutorials - https://www.cgl.ucsf.edu/chimera/tutorials.html
BioPython - https://biopython.org

BepiPred-2.0:
- http://www.cbs.dtu.dk/services/BepiPred/
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5570230/