Lee Xiong An
Personal Project 1
Predicting Cell Penetrating Peptides.

Background

There are three main classes of medicinal drugs with very different sizes (0.1 - 150kda). These drugs are small molecules (<0.9kDa) , peptides ~1kda and large antibodies ~150kda. Small molecules easily enter cells but are less specific due to their small size, while large antibodies are very specific but do not enter cells at all. Peptides are the middle ground being specific yet small enough to enter the cells.

However, the mechanism in which peptides enter the cell are unknown, and under active research. I wish to see if deep learning can find features or patterns explaining why certain peptides enter cell and others do not.

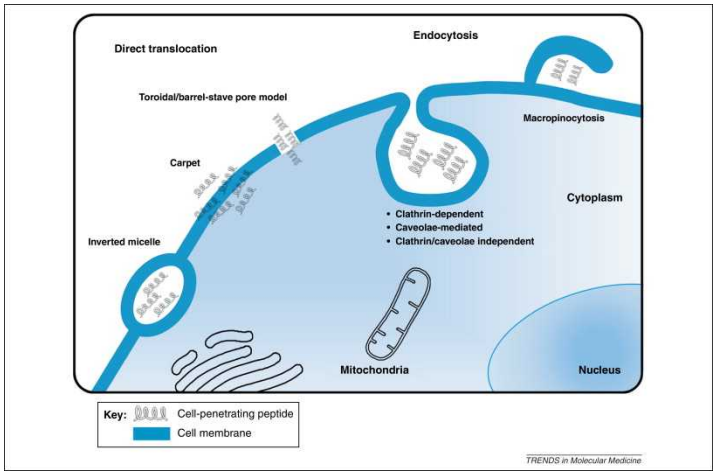

Figure : Cell penetration by peptides via postulated mechanisms.

Data format

Data are present as amino acid sequences. Amino acids contain 20 non-unique characters (ACDEFGHIKLMNPQRSTVWY), and these sequences are usually from 4-30 amino acids long. Furthermore, as amino acids are well characterized experimentally, additional thermodynamic features will be given (4 features). Outputs will be permeable or non-permeable classification problem

Data collection : (what I write because of merck)

Our dataset after preprocessing has 500 positive and 9200 negative sequences.

Since there was a ratio of 1 : 18 positive to negative data-points, I duplicated the positive dataset 18 times for the training set to balance the train set. This improved the AUC metric notwithstanding that AUC is already a balanced metric.

A public dataset containing 100 positive and 34 negative examples will be given due to non-disclosure agreements.

Model :
I used a two models: A) LSTM and B) Convolution NN since these models accept peptide sequences of varying lengths. Both models accept the same inputs and use a similar dense output layer with similar number of parameters ~800k.

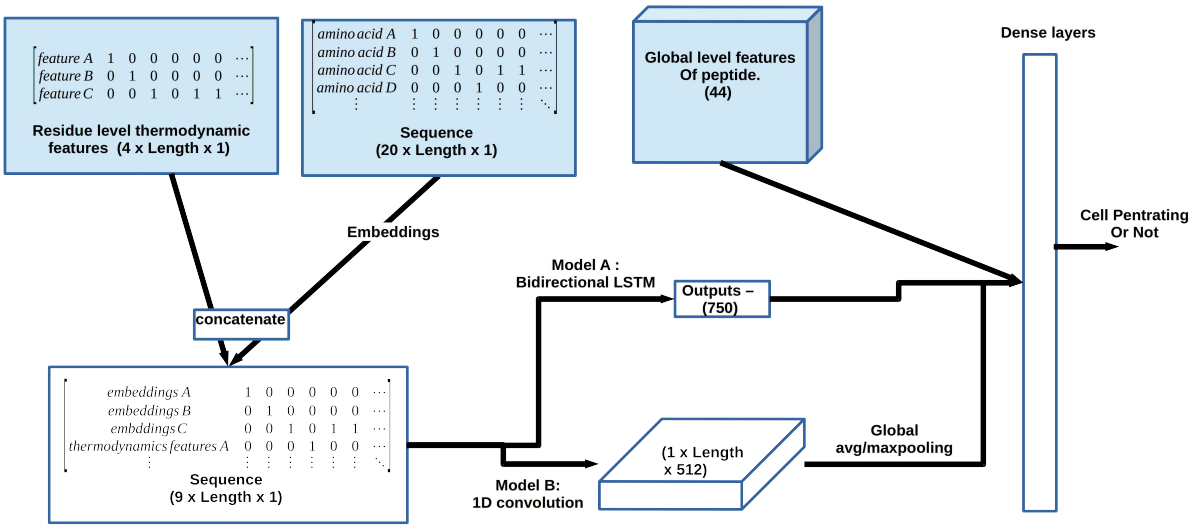

Figure 2 : Inputs colored solid blue, which are residue level thermodynamics features, sequence and global features of peptide. Global features are sequence and thermodynamic features averaged over the entire length of the peptide while residue features are thermodynamics features over each amino acid.

Model training :
The data was split into train, validation and test in ratios 3:1:1 with a total of 20 combinations of train/validation/test. The model was trained with stochastic gradient descent (batch for RNN) to sample the global minima of the cost function. To sample the the local minima, the learning rate has a cosine decay over each epoch. Models were picked based on validation AUC, and the final performance of the model based on the test set was noted down.

Models AUC:
Convolution NN – 0.930
Xgboost – 0.954
LSTM – 0.933
Ensemble model – 90% xgboost, 5% CNN/LSTM – 0.958