# Related Works

* ***BLASTp:***
  + Stephan 1990 [1], synthesize a way of rapid sequence comparison or basic local alignment search tool (BLAST) which directly approximates alignments that optimize a measure of local similarity, the maximal segment pair (MSP) score. BLAST finds regions of similarity between biological sequences. The basic logic behind it is to compare the protein sequences to sequence databases and calculate the statistical significance. BLASTp (Protein BLAST) compares one or more input protein sequences to one or more protein sequences in the database. BLASTp, can be implemented in various ways and can also be applied to different scenarios such as straight-forward DNA and protein sequence database searches, motif searches, gene identification searches, or also in the analysis of multiple regions of similarity in long DNA sequences. In addition to its flexibility and tractability to mathematical analysis, BLAST is an order of magnitude faster than a number of sequence comparison tools of comparable sensitivity.
* ***Profile HMM Models:***
  + Eddy 2011 [2], described an accelerated heuristic known as the ‘‘multiple segment Viterbi’’ (MSV) algorithm for profile Hidden Markov Models (HMMs). Profile HMMs and probabilistic inference methods had already made important contributions in this field. However, practical use of profile HMM methods had been hindered by the computational expense of existing, at the time of 2011, software implementations but the results were still not satisfactory. The way MSV algorithm works is that it computes an optimal sum of multiple ungapped local alignment segments using a striped vector-parallel approach which was previously described and used for fast Smith/Waterman alignment. MSV scores follow the same statistical distribution as gapped optimal local alignment scores, allowing rapid evaluation of significance of an MSV score and thus facilitating its use as a heuristic filter which in turn cuts down on the training time and resources which is required to train the model.

# Proposed Solution:

The project’s main aim is to be able to correctly classify (or predict) an amino acid sequence to the correct family of proteins it belongs to. By using different models on the given dataset (e.g CNN, i.e. ProtCNN and LSTM), we should be able to match the amino acid sequences which are unaligned with the annotations through the collection of 17929 protein families in the Pfam database.

The novel approach that we incorporated into this project is the evolution of the ProtCNN by taking help of Cartesian Genetic Programming. By using predefined building blocks for the CNN model, we constructed a deep learning model by joining the building blocks using a linked list.

# Results:

We used the Pfam Dataset, as a benchmark and compared the results of deep learning models with already existing state-of-the-art alignment based methods such as BLASTp [1] and profile HMM models [2]. The dataset is split in two different ways, the first being the random split of the 17929 families in 80-10-10 partition meaning, 80% training split, 10% validation split and 10% training split. The other split being a clustered split which is done specifically for the distant homologs. Some of the protein sequence may be one of the outliers and random split does not exactly train the model for the outliers and thus, to train the model on the outliers, we performed a clustered split and then trained the models on the clusters.

One thing to keep in mind is that since proteins generally are related through evolution, it causes a wide variety of similarities to occur across different sequences which when split randomly land across different sets and in turn cause similarities to occur over different splits. To eradict this issue we factor in the maximum percent identity of each test sequence with sequences in the train test.

For 30-90% maximum identity with the training set, our model ProtCNN performed significantly better than the alignment-based methods. It was also observed that grouping several ProtCNN and then taking the majority of the predicted outcome also has an effect on the results. Furthermore, it was also observed that although a lot of different groupings affected the results, a group of 13 ProtCNN model works best. It was labelled as ProtENN. ProtENN is significantly more accurate than alignment-based methods for all sequence bins with less than 90% identity to the training set. Using the held-out test error rate as a function of the maximum percent sequence identity across all training sequences we got the results as mentioned in Figure 1 and Figure 2

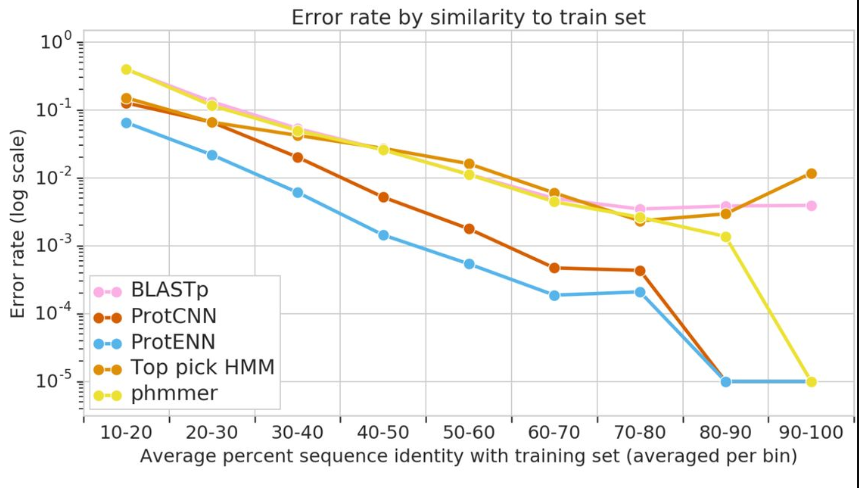


Figure - Random split results

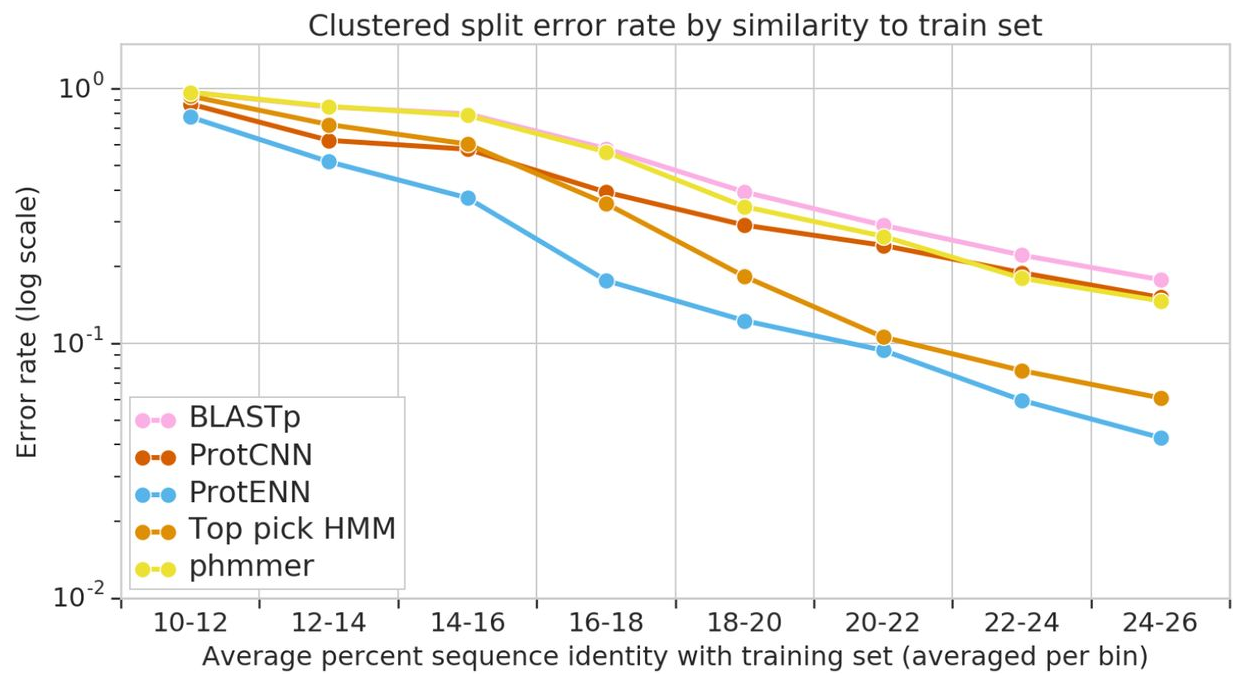


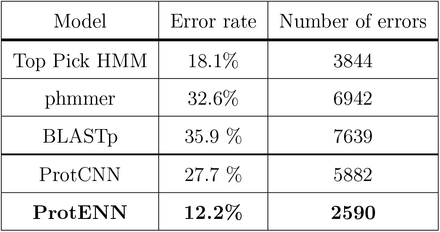
Figure - Clustered Split Results

As you can probably see from Figure 1, both the deep learning models, i.e. ProtCNN and ProtENN make fewer errors than the already existing models such as BLASTp and HMM.

Figure 2 shows that although Top pick HMM did perform better than ProtCNN, it was actually the ProtENN model which was able to deal with the clustered data in the best way possible and emerge victorious when pitched against other methods.

A numerical analysis of the above observation can be seen in the Table 1 below.

Table Error rate of different Models



# Bibliography

|  |  |
| --- | --- |
| [1] | W. G. W. M. E. W. M. a. D. J. L. Stephen F Altschul, "Basic Local Alignment Search Tool," *Journal of molecular biology,,* 1990. |
| [2] | S. R. Eddy, "Accelerated Profile HMM Searches," *PLOS Computational Biology,* 2011. |