*Intuition*:

"With sharp increasing in biological sequences, the traditional sequence alignment methods become unsuitable and infeasible." (https://www.nature.com/articles/s41598-017-12493-2)

This is also our motivation of using deep learning to try to solve this problem.

Since the majority of effects and proteins have less than 10 apparitions, and given the fact that there are over 5800 distinct effects and 17000 proteins, there are too many labels for the data to be clustered in a straight forward way.

This limitations being given, we tried to find a neural network that we can teach the "gene logic", since the input/label size is very small, but the input is big.

If such a "gene logic" exists, meaning that the structure of the codons imply the effect or protein resulted from any gene, than finding some general way of getting a list of relevant features of the gene is feasible.

How? By comparison.

Soft assignation

Intuitive: if aliens would abduct a human and gave him an atlas with 1000 individual animals that were part of 200 species, the best chance of the human to classify a new individual is by making an intuition of what is relevant on that planet for two animals to be part of the same species. For example, on this specific planet, the number of vertebras may be completely irrelevant or even having or not a vertebral column. But the size of the thumb would be a dead giveaway. How would he do this? By comparing individuals of the same species and of different species. In this way, he would observe what particularizes and what doesn't. Now, when he sees a new animal, the human easily identifies what relevant traits align with a particular set of already seen individuals' traits, which probably form a species. We can see how the human both gets an idea of how different two individuals are species-wise, so even if he figures out what a specific species is formed of or simply can tell how distinct two animals are, he can easily cluster the animals and therefore finding species using the difference he learned from his training.

This is where the Siamese LSTM model (more specifically the MaLSTM model) comes into place.

To use this model, we form pairs of genes and say if they have the same protein/effect or not.

When making pairs, "probable" and "unknown" proteins are ignored, since they may ruin the "gene logic". It is not a big sacrifice since there are 103 "unknown" and 2462 "probable" out of 101041 genes.

*Actual method*:

The genes were split in two groups, 96000 for training and testing and 4000 for validation.

The 96000 genes were paired randomly in 15000 pairs, half of the pairs being part of the same cluster, based on the common effects and proteins (as viewed as labels for the genes). The pairs were labeled based on this. So, the training input for the MaLSTM looks like this:

“gene1” (the firs gene sequence), “gene2” (the second gene sequence),”has\_same\_effect” (label of the pair)

Based on the codon2vec representation (linked in the resources), two identical subnets were built of 50 sequential LSTM layers of 80 units. The input was trimmed to 80 codons and it took more than 8 hour to train, the optimal number of codons is more than 500, but we lacked computer power, this has surely affected the result.

The model has an accuracy of 66.42%, which is reasonable.

Now, based on this network, we make pairs of the 96000 genes with a new gene and we look at the first 100 most alike to the new gene. Now, the effects in these 100 genes are counted and the effects are sorted based on how many times they appear. To test how reasonable this method is, the most common 20 effects are compared to the actual effects of the new gene.

The means of the relevant measurements are

* Accuracy:0.297666
* Recall: 0.34294
* Precision: 0.20348

But, if we test for how many of the genes have at least 25% effects discovered, the accuracy becomes 72%.

There is no literature on this method being used in computational biology yet and this gives no clear objective for this method and some improvements can be made (running on 800 codons instead of 80, initial clustering, internal neural network structure), but the core method is implemented and functional;

Some resources:

Siamese LSTM article: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5841810/>

<https://github.com/philippmuench/codon2vec>

<https://github.com/jingcheng-du/Gene2vec> (it’s for the human genome)