**Phylogenetic Tree Assembly**

Itzik Vainsenker 302280144

Oz Granit 203202684

The Problem:

A phylogenetic tree, describes the evolutionary relationship among a set of organisms,

genes, or genomes. Since there is a vast amount of possible trees that can be assembled from a given set of sequences ,All the techniques for inferring the most likely tree have to compromise on either accuracy or runtime.

Goal: Reconstruct the tree which best explains the evolutionary history of this gene.

The Data:

21,152 datasets, each dataset contains ~10-~50 different species.

Each dataset is a set of N aligned sequences (genes, a string made from ‘atcg’ characters of length ~200-~5000) from N different species.

Performance Metrics:

We use log-likelihood analysis of phylogenetic trees as our performance metric, the higher the score, the better the resulting tree.

Likelihood is a common optimization criteria in phylogenetic (Felsenstein 1981)

Baseline Approach:

A standard approach in phylogenetic tree inference is to apply hill-climb heuristic: given the current tree, test all trees that are similar in topology (i.e., neighboring trees) and move to the tree with the highest likelihood.

In the hill-climb approach, our challenge is to determine the method that the next step in the “climb” will be decided by. Essentially, our method will follow the pattern:

* Choose an initial tree
* Extract relevant features from the tree
* Usea **Machine Learning trained model** to predict a set of neighboring trees with the highest likelihood
* Calculate the likelihood of the trees from the predicted set and “climb” to the one with the maximum likelihood
* Repeat

The process ends when none of the set of neighboring trees has a score higher than the current one.

In the description above the definition of a neighbor and the likelihood of a tree remained vague.

The possible neighbors of a tree are chosen by a process called pruning and regrafting (SPR). An SPR neighbor is obtained by pruning a subtree from the main tree and regrafting it to the remaining tree.

We take the SPR neighbor tree space and log-likelihood calculations as given from the biology domain.

Machine Learning Improvements:

For reference we use the baseline method described above, and as a basic machine learning model we use a Random Forest.

We will try to improve results by:

* Usingdifferent Machine Learning models, replacing random forest (for e.g. a NN)

Using deep-learning we can insert as many features as we wish, even a tensor representing the whole tree and have the neural-network learn the right features.

Using that whole tree might not give us good results therefore we thought about implementing a network inspired by CNN with filters that cover nodes and their neighbors, also looking into GNN.

* Choosing an initial tree Usinga Machine Learning model, by which possibly reducing the hill-climb runtime.

Roll-out SPR neighbor:

* Usinga Machine Learning model to predict the best 2 or 3 step neighbor, not just one step neighbor, or even having the neighbor degree per step as a parameter.

Reinforcement learning attempt:

* Train Double Deep Q Network (model-free, off-policy):
  + Current Q-network w is used to select actions
  + Older Q-network w is used to evaluate actions
* End-to-end learning of values Q(s, a) from features s (current tree and clipped data)
* Input state s is stack of raw MSAs (=multiple sequence alignment) and current tree
* Output is Q(s, a) for a finite number of move-type('prune', 'rgft' or 'merged')Xtree\_node positions
* Reward is change in likelihood for that step

We consider a method preferable if it has a better accuracy OR runtime.

Regression-Model:

This part is a description of our attempt to build a linear-NN model to predict the log-likelihood value of a given tree, based on a list of features, replacing the need to calculate said value.

When taking into consideration that the hill-climbing approach requires calculating the log-likelihood value of all SPR neighbors (O(n^2) of tree nodes) of a given tree on-every step of the hill climb, this estimation saves a great deal of calculations.