A machine-learning approach to multiple sequence alignment scoring

Ksenia Polonsky

Under the supervision of Prof. Tal Pupko

Joint work with: Nimrod Serok, Haim Ashkenazy, Itay Mayrose, Jeffrey Thorne

The inference of multiple sequence alignments (MSAs) is at the heart of bioinformatics and genomics. Given a set of biological sequences, MSAs allow the identification of homologous positions across different sequences; thus, each column in an MSA represents characters derived from the common ancestor. Downstream analyses, such as phylogenetic tree inference, ancestral sequence reconstruction, predicting functionally important residues, as well as protein structure and function prediction, were shown to rely heavily on the quality of the inferred MSA. Unlike pairwise alignments, MSA construction of more than a few sequences is computationally demanding. Current algorithms for MSA inference apply different approximate methods. Nearly all approaches currently used to infer MSAs try to find the alignment with the highest "sum-of-pairs" score. This score function is defined as the sum of the pairwise alignment score over all the pairs of sequences in the dataset.

In my research, I aim to develop a better scoring method for MSAs than the sum of pairs. Specifically, I am trying to find an improved score formula using machine learning, i.e., I will ask the computer to find an optimal score. I generated a database of "true" and inferred MSAs to train the machine-learning model. A good scoring function would give a higher score to inferred MSAs that are similar to the true MSAs and a low score to inferred MSAs that are more different from the true MSA. For "true" MSAs, we rely on simulations or benchmarked empirical MSAs generated by considering structural alignments between proteins.

Our preliminary results show that machine learning can be efficiently applied to suggest novel scores that are substantially more accurate than the commonly used sum-of-pairs scores. We show that our machine-learning-based score can help generate more accurate MSAs from a set of unaligned sequences. We expect that in the near future, machine-learning algorithms will replace existing algorithms for aligning many sequences.