Deep learning for extremely fast protein similarity search

Roman Feldbauer[a,b], Arthur Flexer[b] and Thomas Rattei[a]

[a] *Division of Computational Systems Biology (CUBE), University of Vienna*
[b] *Austrian Research Institute for Artificial Intelligence (OFAI)*

Background: Biological sequence analysis sheds light on evolutionary relations, and facilitates knowledge transfer, i.e. obtaining annotations from homologous sequences. Pairwise alignments quantify the similarity between two sequences. The SIMAP2 database contains sequence similarities between all pairs of approximately 30 million proteins. Computation of exact alignments required ten million CPU hours on VSC, and the results necessitate 60 TiB storage capacity.

Problem: Alignment-based similarity search is prohibitively time-consuming for two reasons: (i) Each alignment is expensive (using exact dynamic programming algorithms), and (ii) computation cost increases proportionally with database growth, because each query must be compared against all indexed sequences. Heuristic algorithms like BLAST or MMseqs speed up pairwise comparisons, but do not reduce the computational complexity of database queries. Biological databases are growing perpetually. Consider, for example, the UniProtKB protein database, which has grown from 60 million to 120 million sequences between 2016 and 2018. As a consequence, similarity search puts increasingly heavy burden on computational infrastructure. Methodological improvements are required to handle this massive computational challenges.

Methodology: We propose a twofold approach to address the issue of increasingly expensive protein similarity search: (i) Embedding protein sequences in a vector space enables the usage of distance measures with low computational cost, given that the space's geometry reflects biological properties. Suitable embeddings can be learned with deep networks. Specifically, Siamese or triplet networks can learn vector embeddings, for which distances metrics (for example, Euclidean or Hamming distances) correspond to complex dissimilarity functions. Training such networks for proteins is accelerated tremendously by the precomputed similarity values from SIMAP2. This allows us to train the networks on far more sequences than a previous study on Siamese networks for DNA embeddings did [1]. (ii) Approximate nearest neighbor (NN) structures allow for vector similarity queries in sublinear time. Given the learned vector embeddings from (i), we create a comprehensive database for extremely fast protein similarity search. For this we build upon established techniques such as locality-sensitive hashing, and more recent concepts such as hierarchical navigable small-world graphs [2].

Results: Preliminary results show that the speed of approach (i) for pairwise sequence comparison is competitive with fast heuristics. At the same time, retrieval accuracy is high in combination with approach (ii). That is, the most similar sequence (according to SIMAP2 ground truth) is retrieved from the approximate NN structure in many cases. While not yet ready for use in production, we expect that fine-tuning of the methodology will allow us to alleviate the computational burden of similarity search. Finally, we hope to create a valuable resource for researchers working on proteins in many disciplines of life science.

References

[1] Wei Zheng, Le Yang, Robert J Genco, Jean Wactawski-Wende, Michael Buck, and Yijun Sun, SENSE: Siamese neural network for sequence embedding and alignment-free comparison, Bioinformatics **bty887** (2018).

[2] Yury A. Malkov, and D. A. Yashunin, Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs, arXiv:1603.09320v4 [cs.DS], (2018).