# Fast and Precise Remote Homology Detection

Martin Heusel and Sepp Hochreiter
Institute of Bioinformatics
Johannes Kepler Universität Linz, 4040 Linz, Austria

Biologists require fast and precise methods for homology detection for example to compare whole genomes. The precision of remote homology detection has recently been enhanced by discriminative methods like support vector machines (SVMs) and by using profiles. However the improved precision was bought at the costs of time complexity. SVM approaches are computational expensive because a query has to be compared to all support sequences. Even more time consuming is the query sequence profile generation, e.g. by PSI-BLAST. Also established methods like BLAST, FASTA, BLAT, or SAM suffer from being slowed down by the currently rapid increase of data bases.

To close the gap between speed and precision, we propose a method which achieves state of the art precision but which is faster than BLAST or SAM. Consequently, our method is not based on SVMs, alignment, or profile generation but is based on a radical different approach. We apply a special recurrent network architecture, called "Long Short-Term Memory" (LSTM), which is able to perform "credit assignment through sequences" (CATS). The concept of CATS allows to automatically extract indicative patterns and useful local and global sequence statistics like hydrophobicity, polarity, volume which are all nonlinearly combined. CATS requires a priori learning on a training set which is generated by extending the positive sequence or a class of positive sequences by a BLAST. Training selects models which extremely fast evaluate a query sequence with computational complexity proportional to the query length.

We tested LSTM on a well known SCOP benchmark data set (`http://www.cs.columbia.edu/compbio/svm-pairwise`) and on the PFAM data base for precision and time complexity at remote homology detection. Only tests for LSTM, PSI-BLAST, and SAM 3.5 were feasible (profile or SVMs would have taken hundreds of years). LSTM is both faster and yielded higher precision (measured in the average area under the ROC curve) than PSI-BLAST and SAM as can be seen in Tab. 1.

| method | ROC SCOP | time SCOP | ROC PFAM | ROC rem. PFAM | time PFAM |
|---|---|---|---|---|---|
| PSI-BLAST | 0.764 | 22h | 0.80 | 0.69 | 30h |
| SAM 3.5 | 0.913 | 35h | 0.85 | 0.76 | 1200h |
| LSTM | 0.932 | 190s | 0.88 | 0.79 | 27h |

Table 1: Results of remote homology detection on the SCOP benchmark (SCOP time is the time needed for processing 20,000 new sequences, "ROC rem." considers only the remote homologous sequences.)