SignalP 6.0: A practical application of a protein language model

In the field of Natural Language Processing (NLP), methods for making machines "understand" human languages are developing rapidly. Tasks such as automated translation and text classification are being handled by deep learning methods, e.g. at companies like Google or Meta. Behind the scenes, these tasks are often processed by *language models* trained to predict a masked word from its context.

In the same way, it is possible to train a *protein language model* by masking individual amino acids and make a deep neural network (typically, a Long Short-Term Memory network or a Transformer) predict them from the sequence context. Such a model can then be used to encode amino acids for downstream prediction tasks such as structure, localization, or function. The strength of this approach is that the language model, which is pre-trained on hundreds of millions of unlabeled sequences, contains implicit "knowledge" of protein properties, which can be exploited in downstream tasks with much smaller numbers of labeled sequences.

One recent example of an application is SignalP 6.0, the latest version of the most popular method for predicting signal peptides from amino acid sequence. Using a pre-trained protein language model (ProtBert) has made it possible, for the first time, to distinguish between all five known types of signal peptides in prokaryotes, including the rare Tat/SPII and Sec/SPIII types. In addition, SignalP 6.0 does not need information about the organism of origin, making it well suited for analysis of metagenomic datasets.