DeepLoc: prediction of protein subcellular localization using deep learning

The prediction of eukaryotic protein subcellular localization is a well-studied topic in bioinformatics due to its relevance in proteomics research. Many machine learning methods have been successfully applied in this task, but in most of them, predictions rely on annotation of homologues from knowledge databases. For novel proteins where no annotated homologues exist, and for predicting the effects of sequence variants, it is desirable to have methods for predicting protein properties from sequence information only. Here, we present a prediction algorithm using deep neural networks to predict protein subcellular localization relying only on sequence information. At its core, the prediction model uses a recurrent neural network that processes the entire protein sequence and an attention mechanism identifying protein regions important for the subcellular localization. The model was trained and tested on a protein dataset extracted from one of the latest UniProt releases, in which experimentally annotated proteins follow more stringent criteria than previously. We demonstrate that our model achieves a good accuracy (78% for 10 categories; 92% for membrane-bound or soluble), outperforming current state-of-the-art algorithms, including those relying on homology information. The method is available as a web server at http://www.cbs.dtu.dk/services/DeepLoc. Example code is available at https://github.com/JJAlmagro/subcellular_localization. The dataset is available at http://www.cbs.dtu.dk/services/DeepLoc/data.php. jjalma@dtu.dk.