n-gram analysis of biological sequences in R

n-grams (k-mers) are subvectors of n characters derived from input sequences. Since n-grams capture positional and compositional properties of a sequence, they are widely used in genomics, transcriptomics and proteomics. Despite the continuous interest, there are only few tools tailored for comparative n-gram studies. Furthermore, the volume of n-gram data is usually very large, making its analysis in **R** especially challenging.

The CRAN package *biogram* [Burdukiewicz et al., 2015] facilitates incorporating n-gram data in the **R** workflows. Aside from the efficient extraction and storage of n-grams, the package offers also a novel feature selection method designed specifically for n-gram data. QuiPT (Quick Permutation Test) uses filtering criteria such as information gain (mutual information) to choose the most informative n-grams. To speed up the computation and allow precise estimation of small p-values, QuiPT uses analytically derived distributions instead of a large number of permutations. In addition to this, *biogram* contains tools designed for reducing the dimensionality of the amino acid alphabet [Murphy et al., 2000], further scaling down the feature space.

To illustrate the usage of n-gram data in the analysis of biological sequences, we present two case studies performed solely in **R**. The first, prediction of amyloids, short proteins associated with the number of clinical disorders as Alzheimer's or Creutzfeldt-Jakob's diseases [Fändrich, 2012], employs random forests [Wright and Ziegler, 2015] trained on n-grams. The second, detection of signal peptides orchestrating an extracellular transport of proteins, utilizes more complicated probabilistic framework (Hidden semi-Markov model) based on the n-gram frequency.

References

Michal Burdukiewicz, Piotr Sobczyk, and Chris Lauber. *biogram: analysis of biological sequences using n-grams*. 2015. URL `http://CRAN.R-project.org/package=biogram`. R package version 1.2.

Lynne Reed Murphy, Anders Wallqvist, and Ronald M. Levy. Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Engineering*, 13(3):149–152, March 2000. ISSN 1741-0126, 1741-0134. doi: 10.1093/protein/13.3.149. URL `http://peds.oxfordjournals.org/content/13/3/149`.

Marcus Fändrich. Oligomeric Intermediates in Amyloid Formation: Structure Determination and Mechanisms of Toxicity. *Journal of Molecular Biology*, 421(4–5):427–440, August

2012. ISSN 0022-2836. doi: 10.1016/j.jmb.2012.01.006. URL http://www.sciencedirect.com/science/article/pii/S0022283612000277.

Marvin N. Wright and Andreas Ziegler. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *arXiv:1508.04409 [stat]*, August 2015. URL http://arxiv.org/abs/1508.04409. arXiv: 1508.04409.