eRum 2016
european R users meeting

October 12-14, 2016 in Poznan, Poland

Welcome Address

Welcome! Have a good time!

Contents

Abstracts

N-gram analysis of biological sequences in R

Michał Burdukiewicz[1⋆], Piotr Sobczyk[2], Małgorzata Kotulska[3], Paweł Mackiewicz[4]

1. University of Wrocław, Department of Genomics
2. Wrocław University of Technology, Institute of Mathematics
3. Wrocław University of Technology, Department of Biomedical Engineering
4. University of Wrocław, Department of Genomics

⋆Contact author: michalburdukiewicz@gmail.com

Keywords: n-gram, k-mer, feature selection, proteomics
R packages: biogram, ranger, slam

N-grams (k-mers) are vectors of n characters derived from input sequences, widely used in genomics, transcriptomics and proteomics. Despite the continuous interest in the sequence analysis, there are only a 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* facilitates incorporating n-gram data in the **R** workflows. Aside from the efficient extraction and storage of n-grams, the package offers also a feature selection method designed specifically for this type of data. QuiPT (Quick Permutation Test) uses several filtering criteria such as information gain (mutual information) to choose significant 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, 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, employs random forests 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) but still uses n-gram data for training.

N-gram analysis of biological sequences in R

Michał Burdukiewicz[1⋆], Piotr Sobczyk[2], Małgorzata Kotulska[3], Paweł Mackiewicz[4]

1. University of Wrocław, Department of Genomics
2. Wrocław University of Technology, Institute of Mathematics
3. Wrocław University of Technology, Department of Biomedical Engineering
4. University of Wrocław, Department of Genomics

*Contact author: michalburdukiewicz@gmail.com

Keywords: n-gram, k-mer, feature selection, proteomics
R packages: biogram, ranger, slam

N-grams (k-mers) are vectors of n characters derived from input sequences, widely used in genomics, transcriptomics and proteomics. Despite the continuous interest in the sequence analysis, there are only a 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* facilitates incorporating n-gram data in the **R** workflows. Aside from the efficient extraction and storage of n-grams, the package offers also a feature selection method designed specifically for this type of data. QuiPT (Quick Permutation Test) uses several filtering criteria such as information gain (mutual information) to choose significant 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, 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, employs random forests 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) but still uses n-gram data for training.

N-gram analysis of biological sequences in R

Michał Burdukiewicz[1*], Piotr Sobczyk[2], Małgorzata Kotulska[3], Paweł Mackiewicz[4]

1. University of Wrocław, Department of Genomics
2. Wrocław University of Technology, Institute of Mathematics
3. Wrocław University of Technology, Department of Biomedical Engineering
4. University of Wrocław, Department of Genomics

*Contact author: michalburdukiewicz@gmail.com

Keywords: n-gram, k-mer, feature selection, proteomics
R packages: biogram, ranger, slam

N-grams (k-mers) are vectors of n characters derived from input sequences, widely used in genomics, transcriptomics and proteomics. Despite the continuous interest in the sequence analysis, there are only a 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* facilitates incorporating n-gram data in the **R** workflows. Aside from the efficient extraction and storage of n-grams, the package offers also a feature selection method designed specifically for this type of data. QuiPT (Quick Permutation Test) uses several filtering criteria such as

information gain (mutual information) to choose significant 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, 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, employs random forests 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) but still uses n-gram data for training.