# biogram: n-gram analysis of biological sequences in R

Piotr Sobczyk[1], Chris Lauber[2], Paweł Mackiewicz[3], Michał Burdukiewicz[3]*
*michalburdukiewicz@gmail.com ◇ michbur@github

[1]Wrocław University of Science and Technology, Faculty of Pure and Applied Mathematics, [2]Dresden University of Technology, Bioinformatics Core Unit, [3]University of Wrocław, Department of Genomics

## Introduction

n-grams (k-mers) are vectors of **n** characters derived from input sequences. Originally developed for natural language processing, they are also widely used in genomics, transcriptomics and proteomics. The *biogram* package allows n-gram analysis of biological sequences and accompanies it with unique functionality for generation of simplified amino acid alphabets.

| | P1 | P2 | P3 | P4 | P5 | P6 |
|---|---|---|---|---|---|---|
| S1 | C | T | T | A | G | T |
| S2 | G | A | A | T | A | C |
| S3 | C | C | C | C | A | T |

Sample sequences. S - sequence, P - position.

| | A | C | G | T |
|---|---|---|---|---|
| S1 | 1 | 1 | 1 | 3 |
| S2 | 3 | 1 | 1 | 1 |
| S3 | 1 | 4 | 0 | 1 |

Unigram counts.

| | P1_A | P2_A | P3_A | P4_A | P5_A | P6_A | P1_C | P2_C | P3_C | P4_C | P5_C | P6_C | P1_G |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| S1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| S2 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| S3 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |

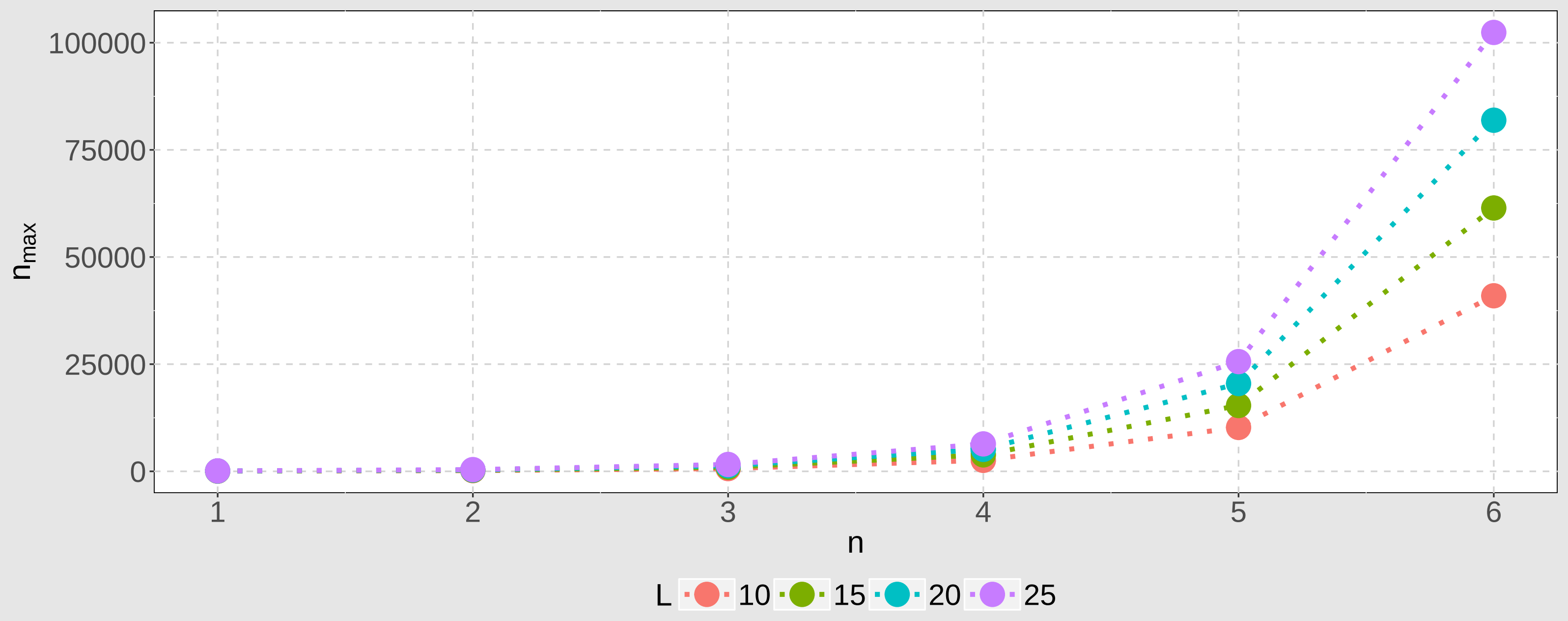
A fraction of possible unigrams with position information.

Function: *count_ngrams()*.

## Curse of dimensionality

Even when we limit ourselves to only continuous positioned n-grams build on $m$ possible characters, feature space growths rapidly with the number of elements in n-gram ($n$) and the length of the sequence ($L$).
The number of possible positioned n-grams:

$$n_{\text{max}} = L \times m^n$$



To decrease $m$, one may reduce the amino acid alphabet using heuristics provided in *biogram*.
Function: *reduce_alphabet()*.

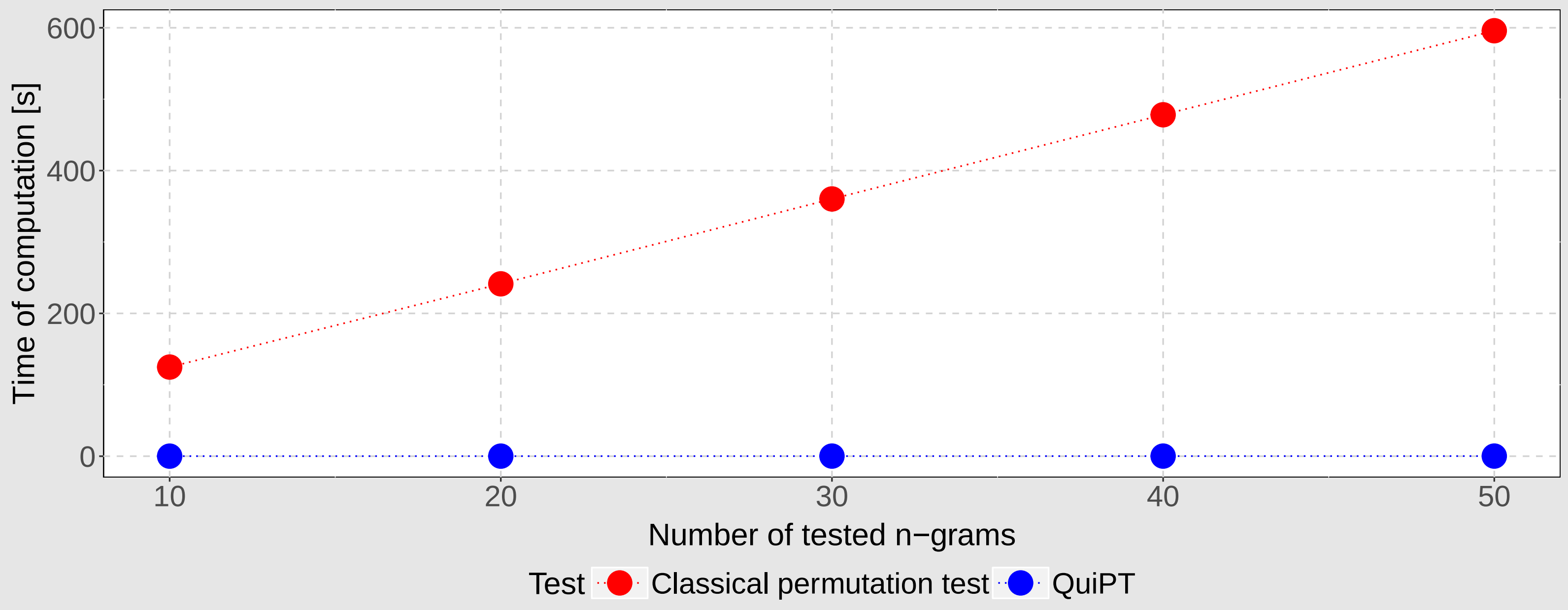
## Selection of important n-grams

Model and statistic independent permutation tests can be used to filter features obtained through counting n-grams.
During a permutation test class labels are randomly exchanged during computation of a significance statistic. p-values are defined as:

$$\text{p-value} = \frac{N_{T_P > T_R}}{N}$$

where $N_{T_P > T_R}$ is number of times when $T_P$ (permuted test statistic) was more extreme than $T_R$ (test statistic for non-permuted data).
Permutation tests are computationally expensive (especially considering precise estimation of small p-values, because the number of permutations is inversely proportional to the interval between p-values).

**Qui**ck **P**ermutation **T**est (QuiPT) thanks to the unique parameterization replaces a permutation test with the exact two-sided Fisher's test reducing the computation cost and returning more precise p-values.



Function: *test_features()*.

## Bibliography

Hiller, K., Grote, A., Scheer, M., Münch, R., and Jahn, D. (2004). PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Research*, 32(suppl 2):W375–W379.

Käll, L., Krogh, A., and Sonnhammer, E. L. L. (2004). A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology*, 338(5):1027–1036.

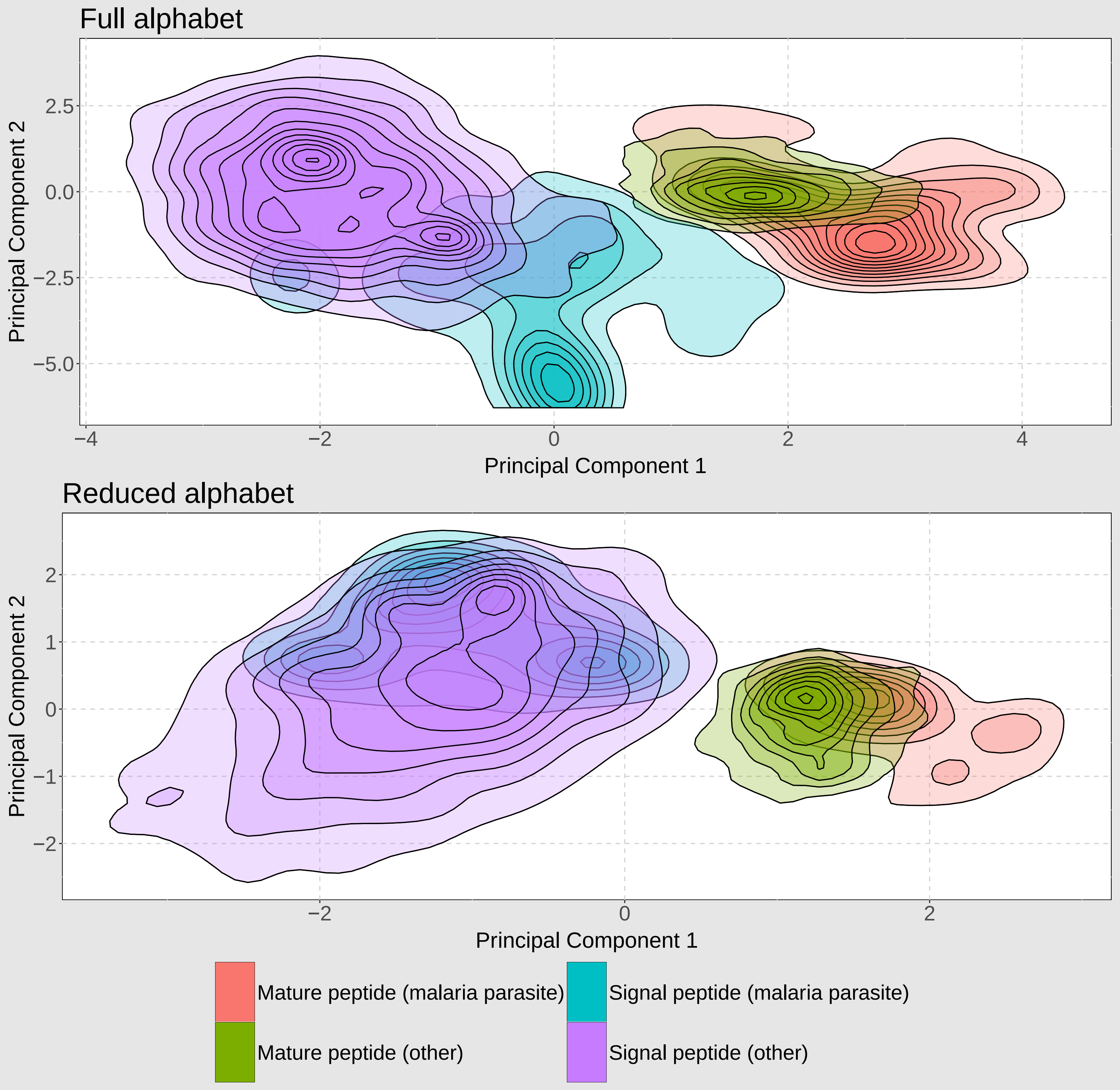
Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*, 8(10):785–786.

Reynolds, S. M., Käll, L., Riffle, M. E., Bilmes, J. A., and Noble, W. S. (2008). Transmembrane topology and signal peptide prediction using dynamic bayesian networks. *PLoS Computational Biology*, 4(11):e1000213.

## Case study: signal peptide prediction

The computational methods for the recognition of signal peptides, short peptides tagging secretory proteins, accurately identify typical peptides, well-represented in protein databases (Petersen et al., 2011). However, these algorithms are not general enough to predict signal peptides with unique amino acid composition, for example those present in proteins from malaria parasites.

## PCA of signal peptides and mature proteins



A countour plot of first two components in Principal Component Analysis of amino acid frequency. The signal peptides from malaria and other taxons differ significantly when the full amino acid alphabet is employed. After the reduction of the alphabet, the signal peptides group together despite their origin.

Here, the reduction of the amino acid alphabet not only creates more manageable feature space, but also mimics the biology behind the process of the signal peptide recognition. Grouping of amino acids reflects their physicochemical properties which are important in protein secretion.

## Benchmark with other predictors of signal peptides

Benchmark data set: 51 proteins with signal peptide and 211 proteins without signal peptide from malaria parasites.
signalHsmm - n-gram based software for prediction of signal peptides.

| | Sensitivity | Specificity | MCC | AUC |
|---|---|---|---|---|
| signalP 4.1 (Petersen et al., 2011) | 0.8235 | 0.9100 | 0.6872 | 0.8667 |
| signalP 4.1 (tm) (Petersen et al., 2011) | 0.6471 | 0.9431 | 0.6196 | 0.7951 |
| PrediSi (Hiller et al., 2004) | 0.3333 | **0.9573** | 0.3849 | 0.6453 |
| Philius (Reynolds et al., 2008) | 0.6078 | 0.9336 | 0.5684 | 0.7707 |
| Phobius (Käll et al., 2004) | 0.6471 | 0.9289 | 0.5895 | 0.7880 |
| signalHsmm | 0.9804 | 0.8720 | 0.7409 | 0.9262 |
| signalHsmm (hom. 50%) | **1.0000** | 0.8768 | **0.7621** | **0.9384** |
| signalHsmm (raw aa) | 0.8431 | 0.9005 | 0.6853 | 0.8718 |

## Conclusions and funding

biogram is a versatile toolkit for n-gram analysis of biological sequences in **R**.

Thanks to the reduction of amino acid alphabet, signalHsmm is able to recognize signal peptides from the malaria parasites and their relatives more accurately than other software.

*biogram* repository: `https://github.com/michbur/biogram`

signalHsmm web-server:
`http://www.smorfland.uni.wroc.pl/shiny/signalHsmm/`.

Find us online: `https://github.com/michbur/USER2017`.

This research was partially funded by the KNOW Consortium and National Science Center (2015/17/N/NZ2/01845).