

# n-gram analysis of biological sequences in R

---

Michał Burdukiewicz<sup>1</sup>, Piotr Sobczyk<sup>2</sup>, Małgorzata Kotulska<sup>3</sup>, Paweł Mackiewicz<sup>1</sup>

<sup>1</sup>University of Wrocław, Department of Genomics,

<sup>2</sup>Wrocław University of Science and Technology, Faculty of Pure and Applied Mathematics,

<sup>3</sup>Wrocław University of Science and Technology, Department of Biomedical Engineering

# Introduction

---

*biogram*: the **R** package for the n-gram analysis of biological sequences.

biogram workflow:

1. Extract n-grams.
2. Change an alphabet.
3. Filter n-grams.

# Biological sequences

Long chains of amino acids (proteins) or nucleotides (RNA or DNA).

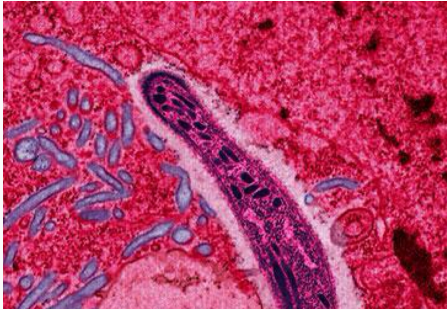
Sample protein sequence:

MRKLYCVLLLSAFEFTYMINFGRGQNYWEHPYQKSDVYHP

INEHREHPKEYQYPLHQEHTYQQEDSGEDENTLQHAYPID

HEGAEPAPQEQNLFSSIEIV...

# Biological sequences



*Plasmodium falciparum*. Source: <http://www.protists.ensembl.org>

Protein of *Plasmodium falciparum*:

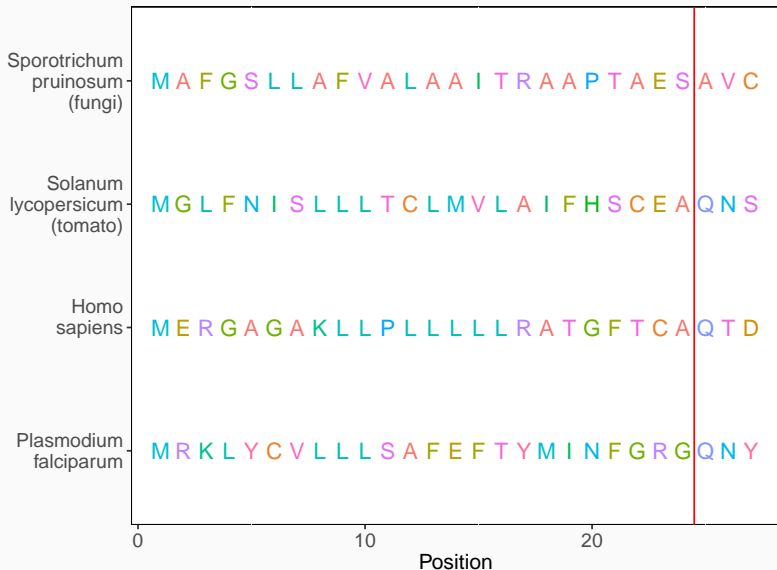
```
MRKLYCVLLLSAFETYMINFGRGQNYWEHPYQKSDVYHP  
INEHREHPKEYQYPLHQEHTYQQEDSGEDENTLQHAYPID  
HEGAEPAPQEQNLFSSIEIV...
```

# Biological sequences

Signal peptide (red): n-terminal amino acid sequence directing proteins to the endomembrane system and next to extracellular localizations.

MRKLYCVLLLSAFEFTYMINFGRGQNYWEHPYQKSDVYHP  
INEHREHPKEYQYPLHQEHTYQQEDSGEDENTLQHAYPID  
HEGAEPAPQEQNLFSSIEIV...

# Biological sequences



## n-grams

---



# n-grams

n-grams (k-tuples) are vectors of  $n$  characters derived from input sequence(s).

	P1	P2	P3	P4	P5	P6	P7	P8	P9
PF	M	R	K	L	Y	C	V	L	L
HS	M	G	L	F	N	I	S	L	L
SL	M	A	F	G	S	L	L	A	F
SP	M	E	R	G	A	G	A	K	L

1-grams: M, M, M, M, R, G, A, E

## n-grams

	P1	P2	P3	P4	P5	P6	P7	P8	P9
PF	M	R	K	L	Y	C	V	L	L
HS	M	G	L	F	N	I	S	L	L
SL	M	A	F	G	S	L	L	A	F
SP	M	E	R	G	A	G	A	K	L

2-grams: MR, MG, MA, ME, RK, GL, AF, ER

## n-grams

	P1	P2	P3	P4	P5	P6	P7	P8	P9
PF	M	R	K	L	Y	C	V	L	L
HS	M	G	L	F	N	I	S	L	L
SL	M	A	F	G	S	L	L	A	F
SP	M	E	R	G	A	G	A	K	L

3-grams: MRK, MGL, MAF, MER, RKL, GLF, AFG,  
ERG

## n-grams

	P1	P2	P3	P4	P5	P6	P7	P8	P9
PF	M	R	K	L	Y	C	V	L	L
HS	M	G	L	F	N	I	S	L	L
SL	M	A	F	G	S	L	L	A	F
SP	M	E	R	G	A	G	A	K	L

2-grams (with a single gap): M-K, M-L, M-F, M-R, R-L,  
G-F, A-G, E-G

## n-grams

	P1	P2	P3	P4	P5	P6	P7	P8	P9
PF	M	R	K	L	Y	C	V	L	L
HS	M	G	L	F	N	I	S	L	L
SL	M	A	F	G	S	L	L	A	F
SP	M	E	R	G	A	G	A	K	L

3-grams (with gaps): M - K - - C , M - L - - I , M - F - - L ,  
M - R - - G , R - L - - V , G - F - - S , A - G - - L , E - G - - A

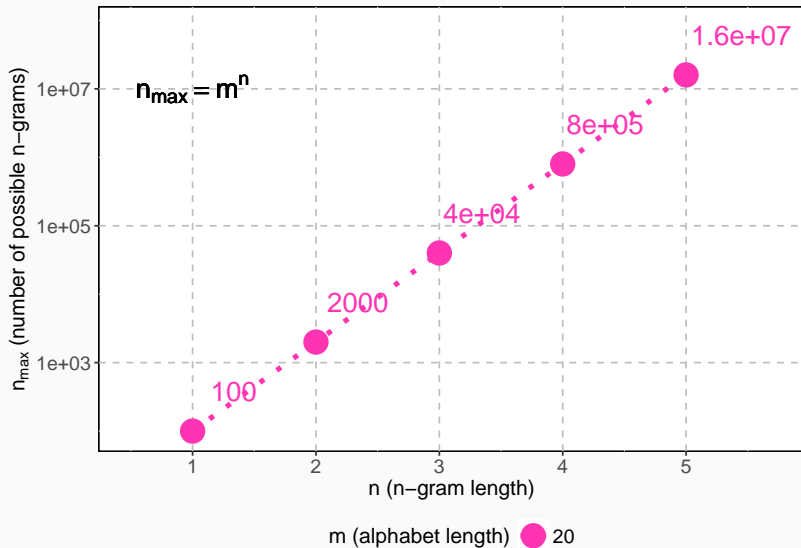
## Proposed model

$$y \sim \text{n-gram}_1 + \text{n-gram}_2 + \dots + \text{n-gram}_n$$

Problems:

- large number of possible n-grams,
- the majority of n-grams is noninformative.

## n-gram counts



## n-gram counts

	A_0	C_0	D_0	E_0	F_0	...	L.F_0	M.F_0	N.F_0	...
PF	1	1	0	1	3	...	0	0	1	...
HS	2	2	0	1	2	...	1	0	0	...
SL	9	1	0	1	2	...	0	0	0	...
SP	4	1	1	1	1	...	0	0	0	...

The sparsity of the n-gram count matrix grows with  $n$ .



# Sparse matrix representation

---

Packages:

- *Matrix*: S4, part of the base **R** distribution (499 depending or importing CRAN packages).
- *SparseM*: S4 (26 depending or importing CRAN packages).
- *slam*: S3, very small matrices (50 depending or importing CRAN packages).

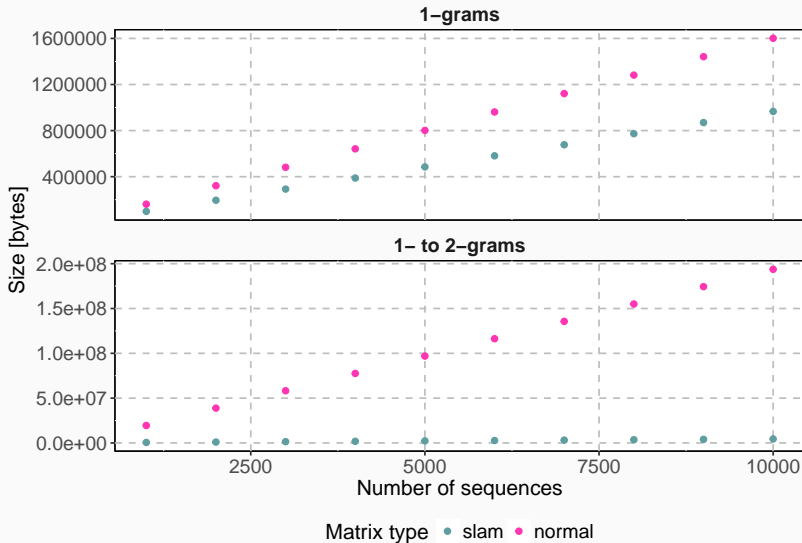
All packages add their methods to common functions and operators for dense matrices.

Packages:

- Matrix: S4, part of base R distribution, well-developed (499 depending or importing CRAN packages).
- SparseM: S4 (26 depending or importing CRAN packages).
- slam: S3, very small matrices (50 depending or importing CRAN packages, including *biogram*).

All packages add their methods to common functions and operators for dense matrices.

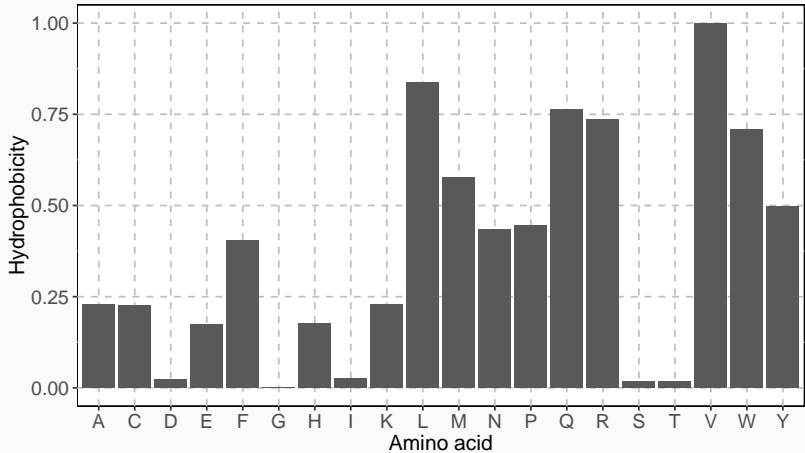
# slam representation



## **Reduction of the amino acid alphabet**

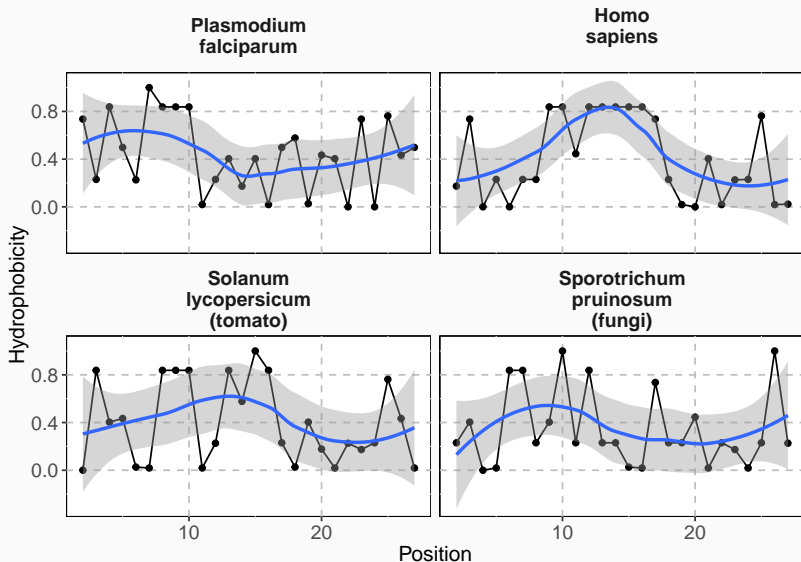
---

# Properties of amino acids

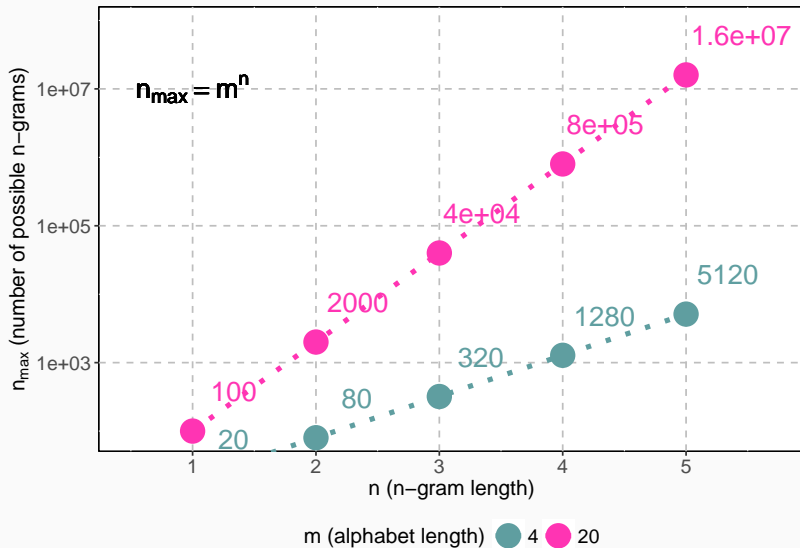


Amino acids may be described using their physicochemical properties.

# Properties of amino acids

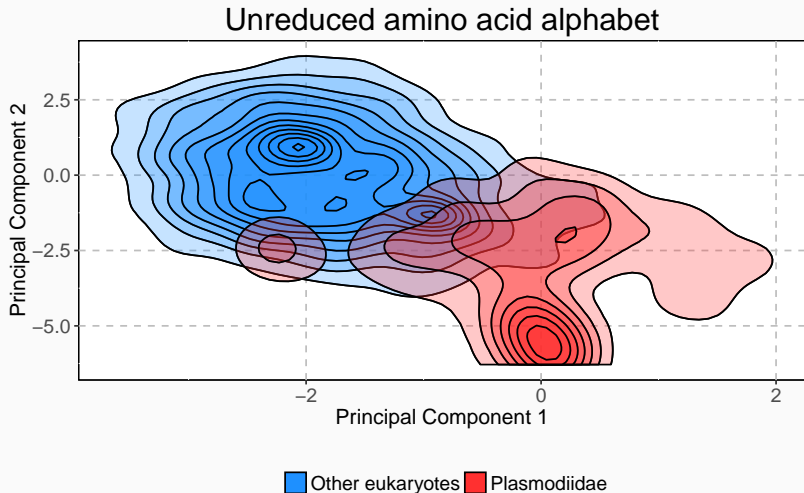


# Reduction of the alphabet





## Reduction of the alphabet

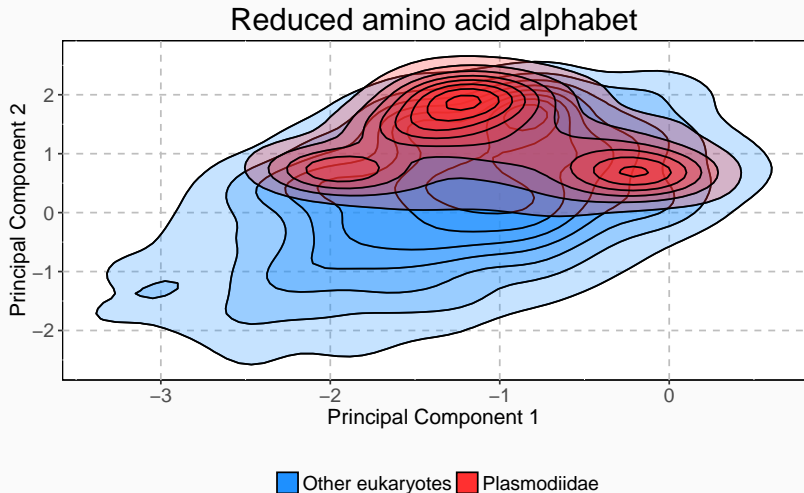


PCA analysis of amino acid frequency in signal peptides.

## Reduction of the alphabet

Group	Amino acids
I	D, E, H, K, N, Q, R
II	G, P, S, T, Y
III	F, I, L, M, V, W
IV	A, C

## Reduction of the alphabet



PCA analysis of amino acid frequency in signal peptides.

## Filtering n-grams

---

# Permutation Test

Informative n-grams are usually selected using permutation tests.

During a permutation test we shuffle randomly class labels and compute a defined statistic (e.g. information gain). Values of the statistic for permuted data are compared with the value of statistic for original data.

# Permutation Test

target	Original data	Permuted data 1	Permuted data 2	...
0	0	1	1	...
0	1	0	1	...
0	1	1	1	...
1	0	0	0	...
1	1	1	0	...
1	0	0	0	...

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

$N_{T_P > T_R}$ : number of cases, where  $T_P$  (permuted test statistic) has more extreme values than  $T_R$  (test statistic for original data).

$N$ : number of permutations.

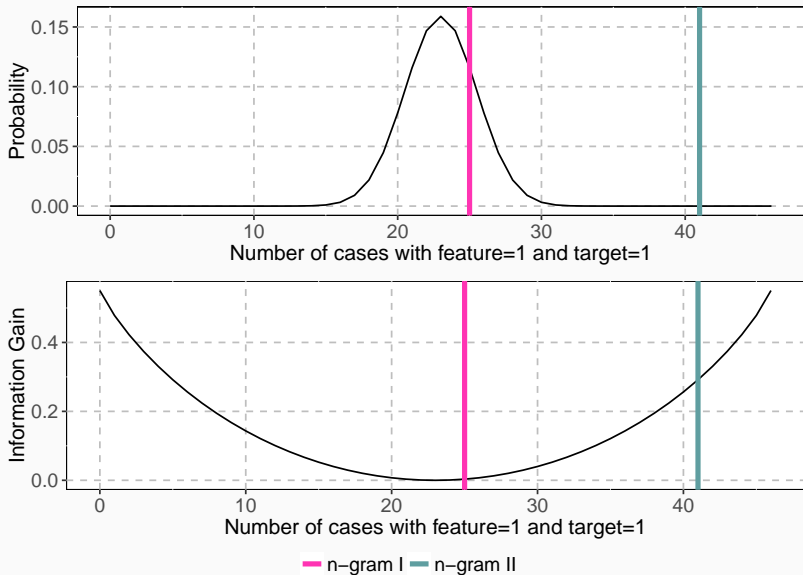
**Quick Permutation Test** is a fast alternative to permutation tests for n-gram data. It computes a probability for a given contingency table providing the exact p-value for the specific value level of the test statistic.

IG: 0.0032

Target	n-gram I	Count
0	0	29
1	0	25
0	1	21
1	1	25

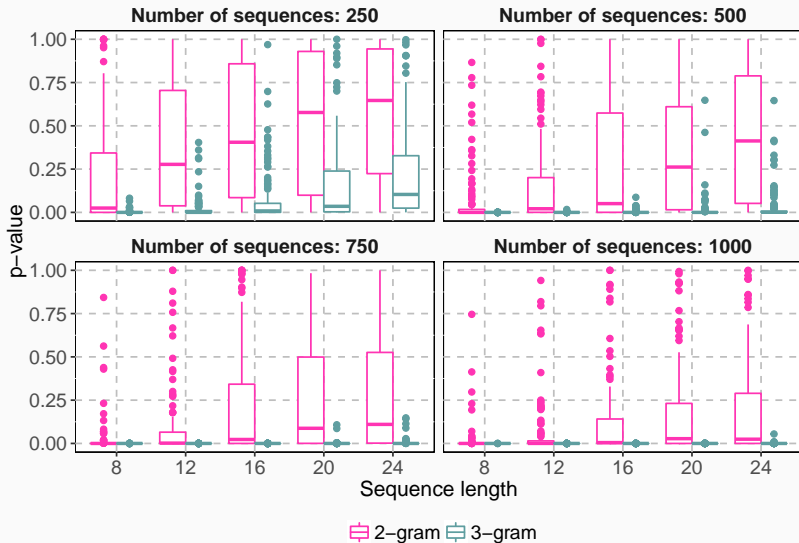
IG: 0.2917

Target	n-gram II	Count
0	0	45
1	0	9
0	1	5
1	1	41





# QuiPT performance



## Summary and conclusion

---

*biogram*: the **R** package for the n-gram analysis of biological sequences.

biogram workflow:

1. Extract n-grams (`count_ngrams()` and `count_multigrams()`).
2. Change an alphabet (`reduce_alphabet()` and `calc_ed()`).
3. Filter n-grams (`test_features()`).

*biogram*: the **R** package for the n-gram analysis of biological sequences.

<https://CRAN.R-project.org/package=biogram> (1.3)

<http://github.com/michbur/biogram> (1.4)

Burdukiewicz M, Sobczyk P, Rödiger S, Duda-Madej A, Mackiewicz P, Kotulska M. (2016) **Prediction of amyloidogenicity based on the n-gram analysis.** *PeerJ Preprints* 4:e2390v1

<https://doi.org/10.7287/peerj.preprints.2390v1>

A novel method of detecting amyloids, proteins involved in many neurodegenerative disorders, such as Alzheimer's or Creutzfeldt-Jakob's diseases based on n-grams used to train a random forest classifier (from the *ranger* package).

## Acknowledgements and funding

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

- Paweł Mackiewicz,
- Małgorzata Kotulska,
- **biogram** package  
(<https://cran.r-project.org/package=biogram>):
  - Piotr Sobczyk,
  - Chris Lauber.