n-gram analysis of biological sequences in R

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

[1]University of Wrocław, Department of Genomics,
[2]Wrocław University of Science and Technology, Faculty of Pure and Applied Mathematics,
[3]Wrocław University of Science and Technology, Department of Biomedical Engineering

Introduction

Biological sequences

Chains of amino acids (proteins) or nucleotides (RNA or DNA).

Sample protein sequence:

```
MKLLLLLLIVSASMLIESLVNADGYIKRRDGCKVACLIGNE

GCDKECKAYGGSYGYCWTWGLACWCEGLPDDKTWKSETNT

CGGKK
```

Biological sequences

Mesobuthus martensii. Source: `http://www.sciencenews.org`

Toxin produced by *Mesobuthus martensii*:

MKLLLLLIVSASMLIESLVNADGYIKRRDGCKVACLIGNE

GCDKECKAYGGSYGYCWTWGLACWCEGLPDDKTWKSETNT

CGGKK

Biological sequences

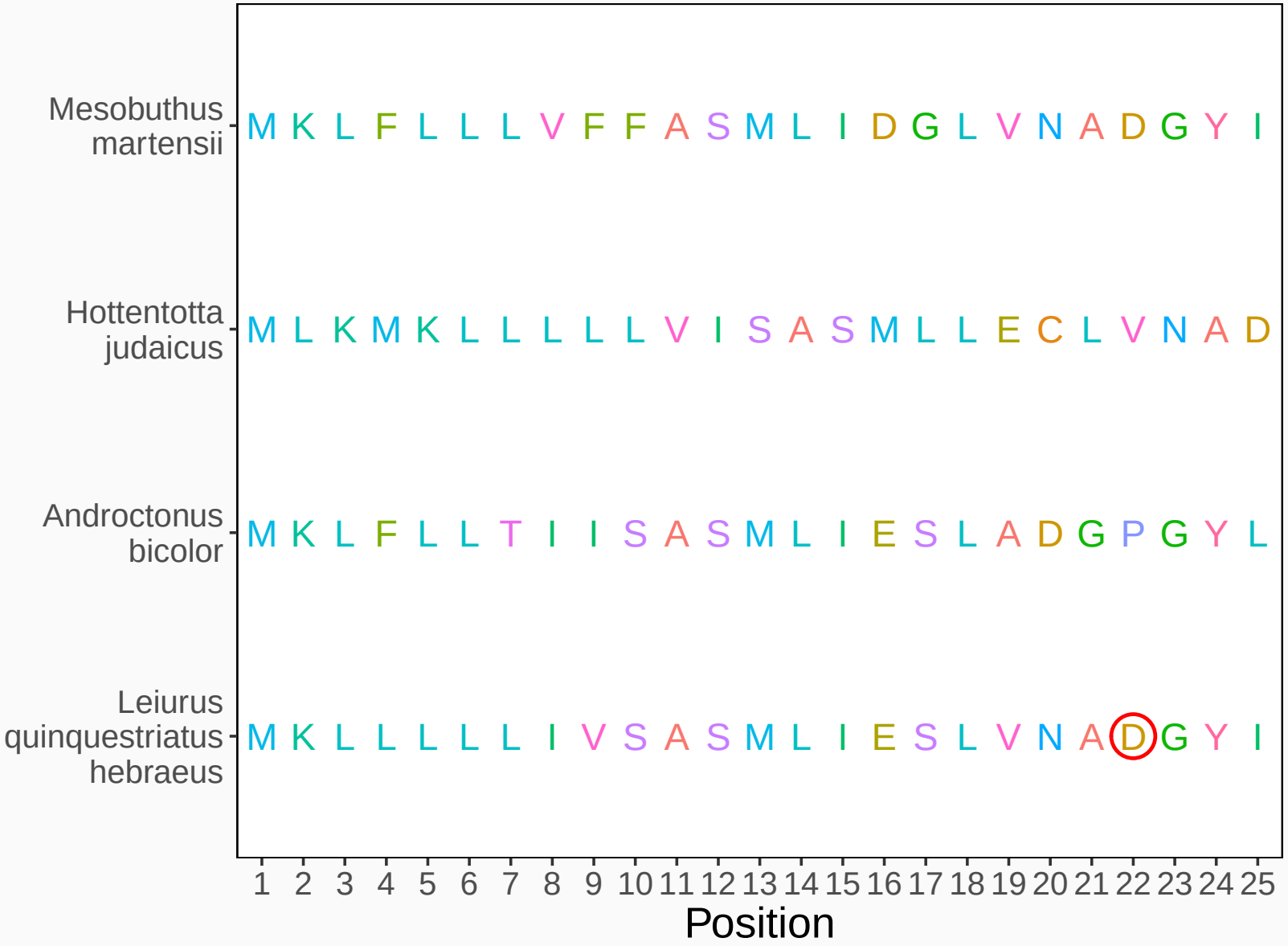
Signal peptide (red): n-terminal amino acid sequence directing proteins to the endomembrane system and next to extracellular localizations.

MKLLLLLIVSASMLIESLVNADGYIKRRDGCKVACLIGNE

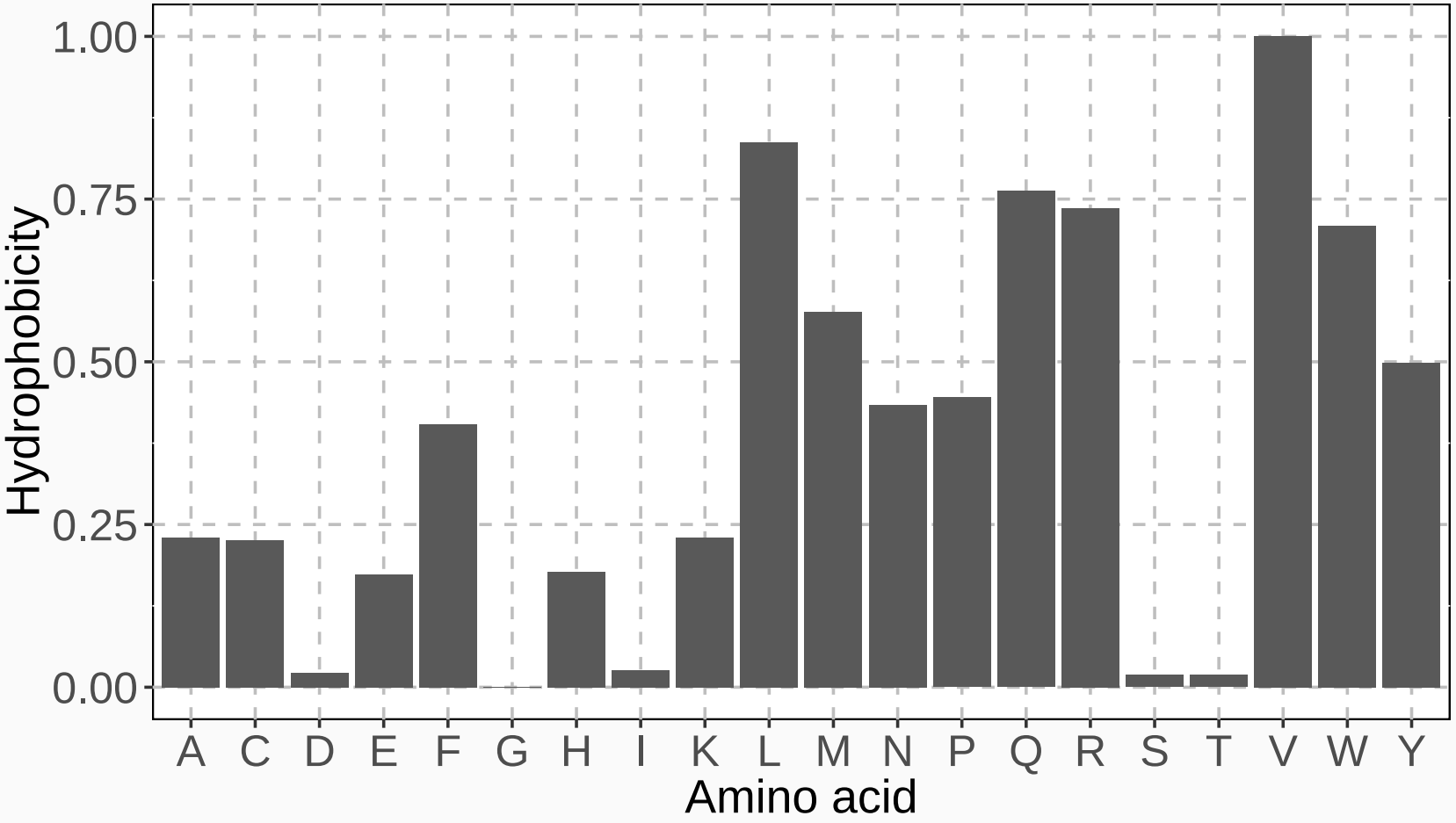
GCDKECKAYGGSYGYCWTWGLACWCEGLPDDKTWKSETNT

CGGKK

Biological sequences



Properites of amino acids



Amino acids may be described using their physicochemical properites.

Biological sequences

n-grams

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