Simplified alphabets in protein analysis

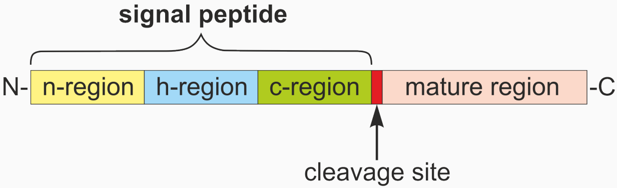
Michał Burdukiewicz

Department of Genomics, University of Wrocław

Outline

Simplified alphabets

Signal peptides

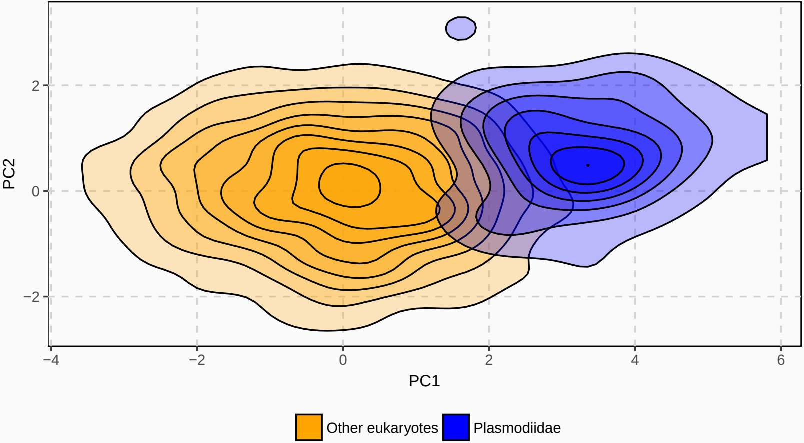


Signal peptides possess three distinct domains with variable length and characteristic amino acid composition (Hegde and Bernstein, 2006):

- n-region: mostly basic residues (Nielsen and Krogh, 1998),
- h-region: strongly hydrophobic residues (Nielsen and Krogh, 1998),
- c-region: a few polar, uncharged residues.

Signal peptides

Amino acid composition of signal peptides differ between Plasmodium sp. and other eukaryotes. Therefore, predictors of signal peptides do not detect malarial signal peptides accurately.



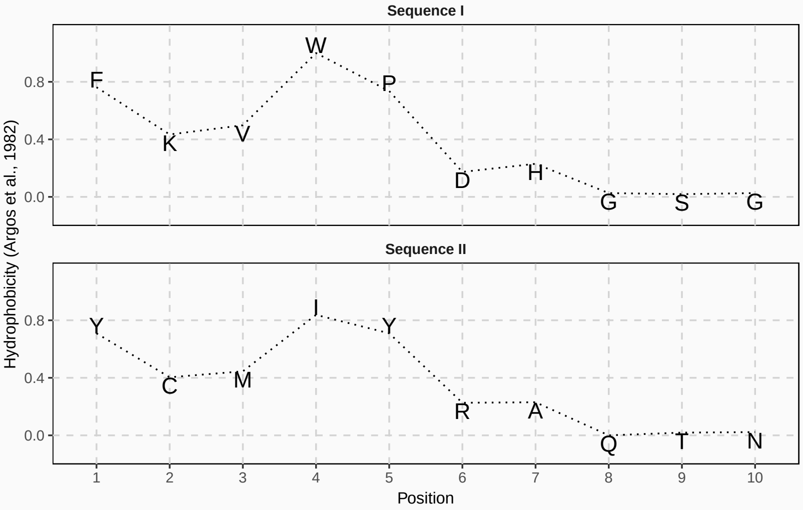
Simplified alphabets

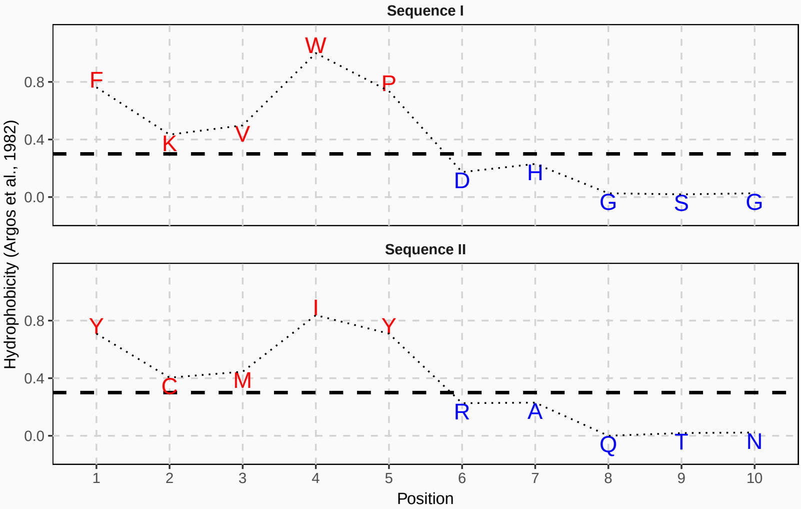
Simplified alphabets:

- are based on grouping amino acids with similar physicochemical properties,
- ease computational analysis of a sequence (Murphy et al., 2000),
- create more explicite models.

Two sequences that are drastically different considering their amino acids composition can have the same physicochemical properties.

Sequence I:	`FKVWPDHGSG`
Sequence II:	`YCMIYRAQTN`


Sequence I
Sequence II
Hydrophobicity (Argos et al., 1982)
0.8
0.4
0.0
F
K
V
W
P
D
H
G
S
G
Y
C
M
I
Y
R
A
Q
T
N
1
2
3
4
5
6
7
8
9
10
Position


Sequence I
Sequence II
Hydrophobicity (Argos et al., 1982)
0.8
0.4
0.0
F
K
V
W
P
D
H
G
S
G
Y
C
M
I
Y
R
A
Q
T
N
1
2
3
4
5
6
7
8
9
10
Position

Subgroup	Amino acid
1	C, I, L, K, M, F, P, W, Y, V
2	A, D, E, G, H, N, Q, R, S, T

Sequence I:	`FKVWPDHGSG`
Sequence II:	`YCMIYRAQTN`

Subgroup	Amino acid
1	C, I, L, K, M, F, P, W, Y, V
2	A, D, E, G, H, N, Q, R, S, T

Sequence I: `FKVWPDHGSG`

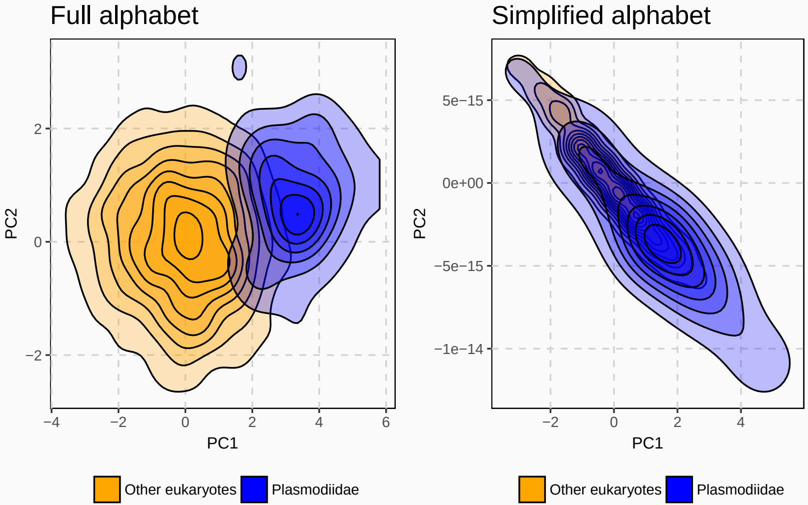
Sequence II: `YCMIYRAQTN`

The best-performing simplified alphabet

Subgroup ID	Amino acids
1	G
2	K, P, R
3	I, L, V
4	F, W, Y
5	A, C, H, M
6	D, E, N, Q, S, T

Group 2 - charged breakers of β-structures.

Signal peptide prediction

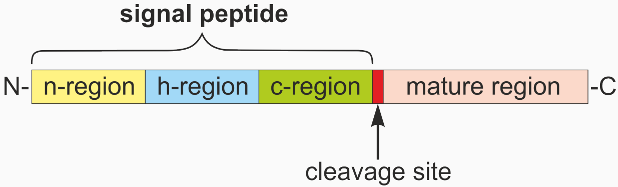


Full alphabet

Simplified alphabet

PCA of amino acid frequency in signal peptides.

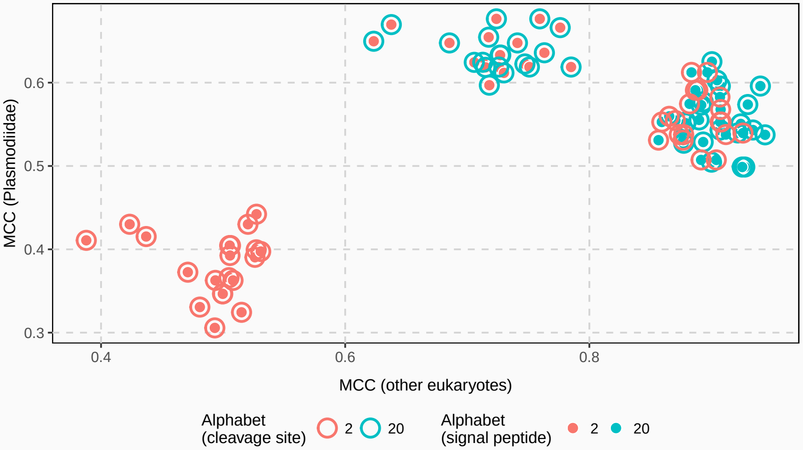
Signal peptide prediction



SignalP 4.1 (Petersen et al., 2011) combines output of two separate predictors:

- cleavage site,
- signal peptide.

Signal peptide prediction



References I

References

Hegde, R. S. and Bernstein, H. D. (2006). The surprising complexity of signal sequences. *Trends in Biochemical Sciences*, 31(10):563–571.

Murphy, L. R., Wallqvist, A., and Levy, R. M. (2000). Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Engineering*, 13(3):149–152.

References II

Nielsen, H. and Krogh, A. (1998). Prediction of signal peptides and signal anchors by a hidden Markov model. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 6:122–130.

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.