signalHsmm: prediction of malarial signal peptides

Michał Burdukiewicz[1]*, Piotr Sobczyk[2], Paweł Błażej[1], Paweł Mackiewicz[1]

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

Signal peptides

Signal peptides:

- are short (20-30 residues) N-terminal amino acid sequences forming α-helices,

Signal peptides:

- are short (20-30 residues) N-terminal amino acid sequences forming α-helices,
- direct proteins to the endomembrane system and next to extra- or intracellular localizations,

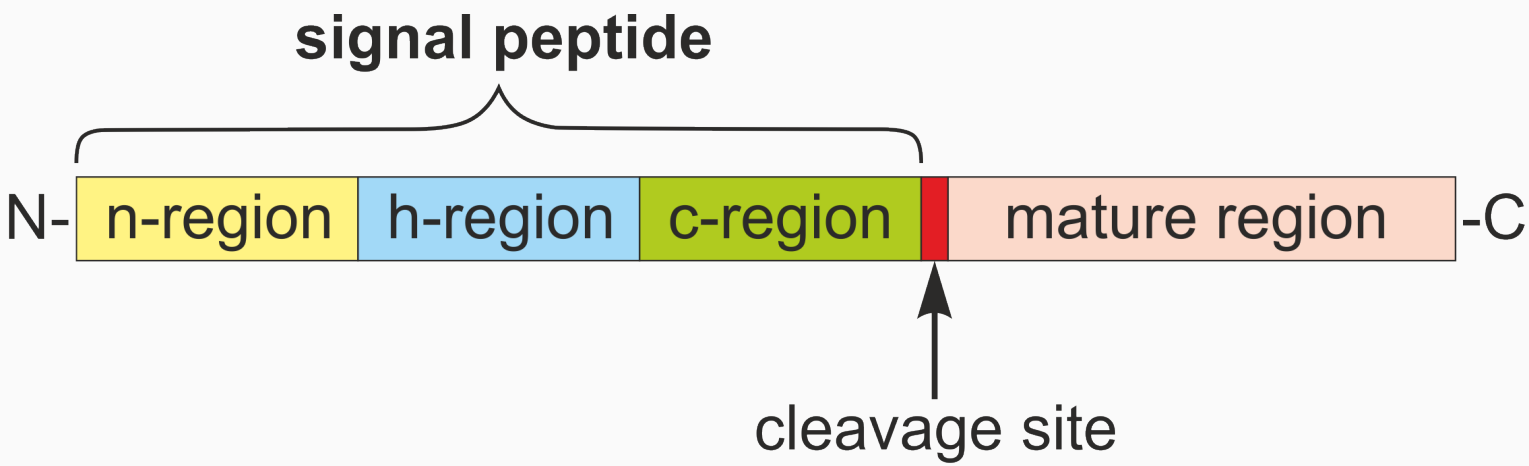
Signal peptides:

- are short (20-30 residues) N-terminal amino acid sequences forming α-helices,
- direct proteins to the endomembrane system and next to extra- or intracellular localizations,
- are universal enough to direct properly proteins in different secretory systems; artifically introduced bacterial signal peptides can guide proteins in mammals (Nagano and Masuda, 2014) and plants (Moeller et al., 2009),

Signal peptides:

- are short (20-30 residues) N-terminal amino acid sequences forming α-helices,
- direct proteins to the endomembrane system and next to extra- or intracellular localizations,
- are universal enough to direct properly proteins in different secretory systems; artifically introduced bacterial signal peptides can guide proteins in mammals (Nagano and Masuda, 2014) and plants (Moeller et al., 2009),
- tag hormons, immune system proteins, structural proteins, and metabolic enzymes.

Architecture



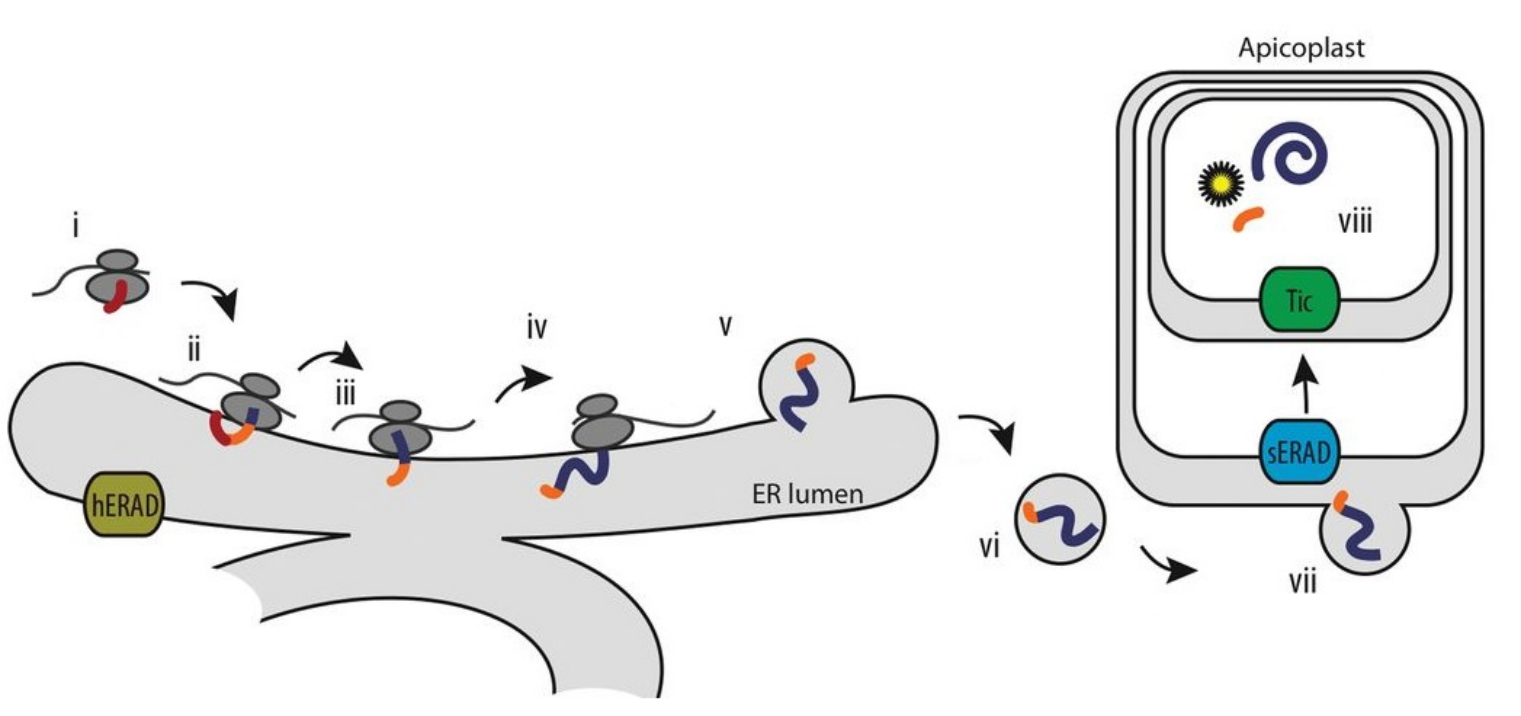
Signal peptides possess three distinct domains with variable length and characteristic amino acid composition (Hegde and Bernstein, 2006):

- 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 (Jain et al., 1994).

Malarial signal peptides

Transit signal

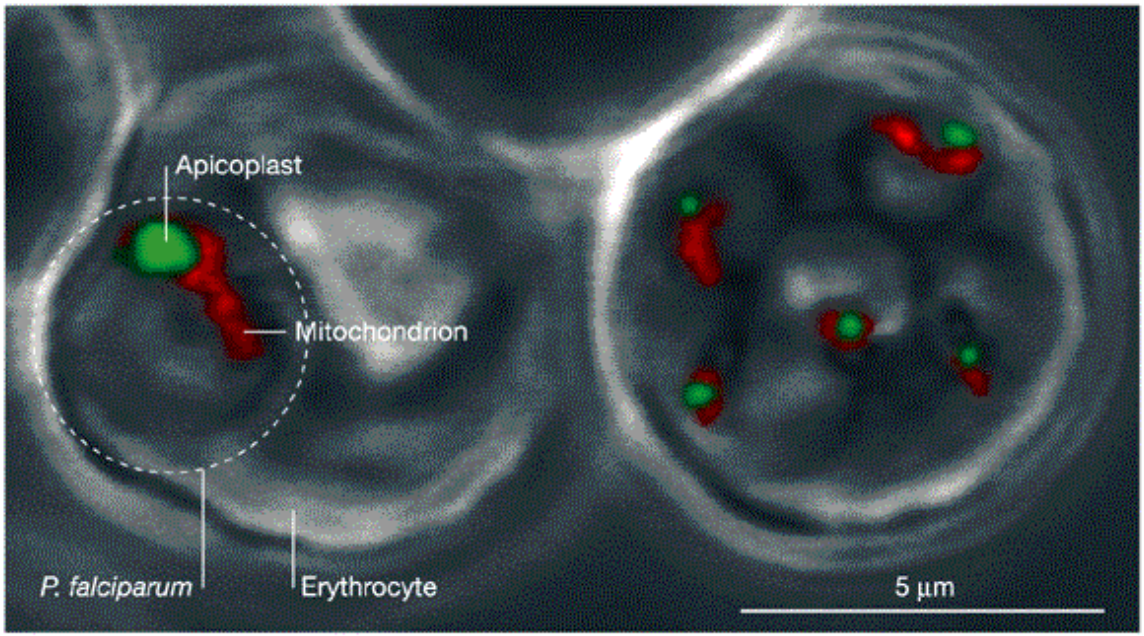
The **signal peptide** is required for targeting proteins to intracellular locations through the endomembrane system, for example Plasmodium-specific apicoplast.



Kalanon and McFadden (2010)

Apicoplast

Four membrane-bounded plastid of Plasmodium sp. responsible for several biochemical pathways including the biosynthesis of fatty acids, isoprenoids and haem.



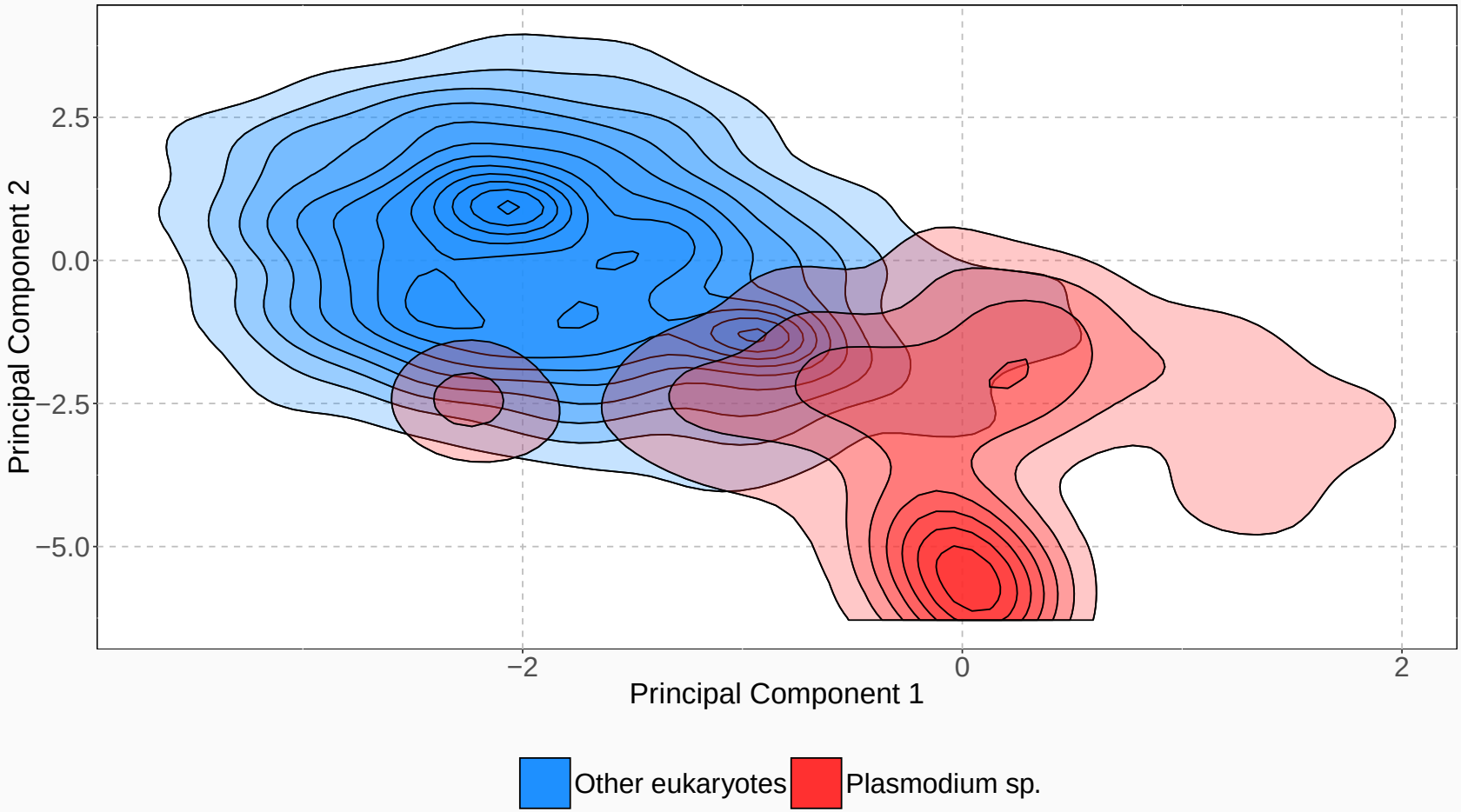
Ralph et al. (2004)

Apicoplast

The absence of a metabolic counterpart in human host make the apicoplast proteins promising targets for anti-malarial drug development.

PCA of amino acid frequency

Heavy adenine-thymine bias of malarial genomes alters amino acid composition of malarial signal peptides making them hard to predict using software trained on other eukaryotes.



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

There are no enough malarial signal peptides to train a specialized predictor.

Aim

Can we employ decision rules used for prediction of eukaryotic signal peptides to correctly detect malarial signal peptides?

Even nonbiological sequences can be effective signal peptides provided they fulfill general requirements (Tonkin et al., 2008).

NH_2-SKINNYSLINKYKINKYTHING-COOH - targets apicoplast.

NH_2-ITWILLNEVERTARGETPLASTID-COOH - does not target apicoplast.

Methods

Reduced amino acid alphabets

To date, several reduced amino acid alphabets have been proposed, which have been applied to (among others) protein folding and protein structure prediction.

Novel reduced amino acid alphabets

13 physicochemical properties handpicked from AAIndex database relevant to the regional architecture of signal peptides.

Property name	Amino acid scale
Size	Size
Size	Molecular weight
Size	Residue volume
Size	Bulkiness
Hydrophobicity	Normalized hydrophobicity scales for α-proteins
Hydrophobicity	Consensus normalized hydrophobicity scale
Hydrophobicity	Hydropathy index
Hydrophobicity	Surrounding hydrophobicity in α-helix
Polarity	Polarity
Polarity	Mean polarity
Occurrence in α-helices	Signal sequence helical potential
Occurrence in α-helices	Normalized frequency of N-terminal helix
Occurrence in α-helices	Relative frequency in α-helix

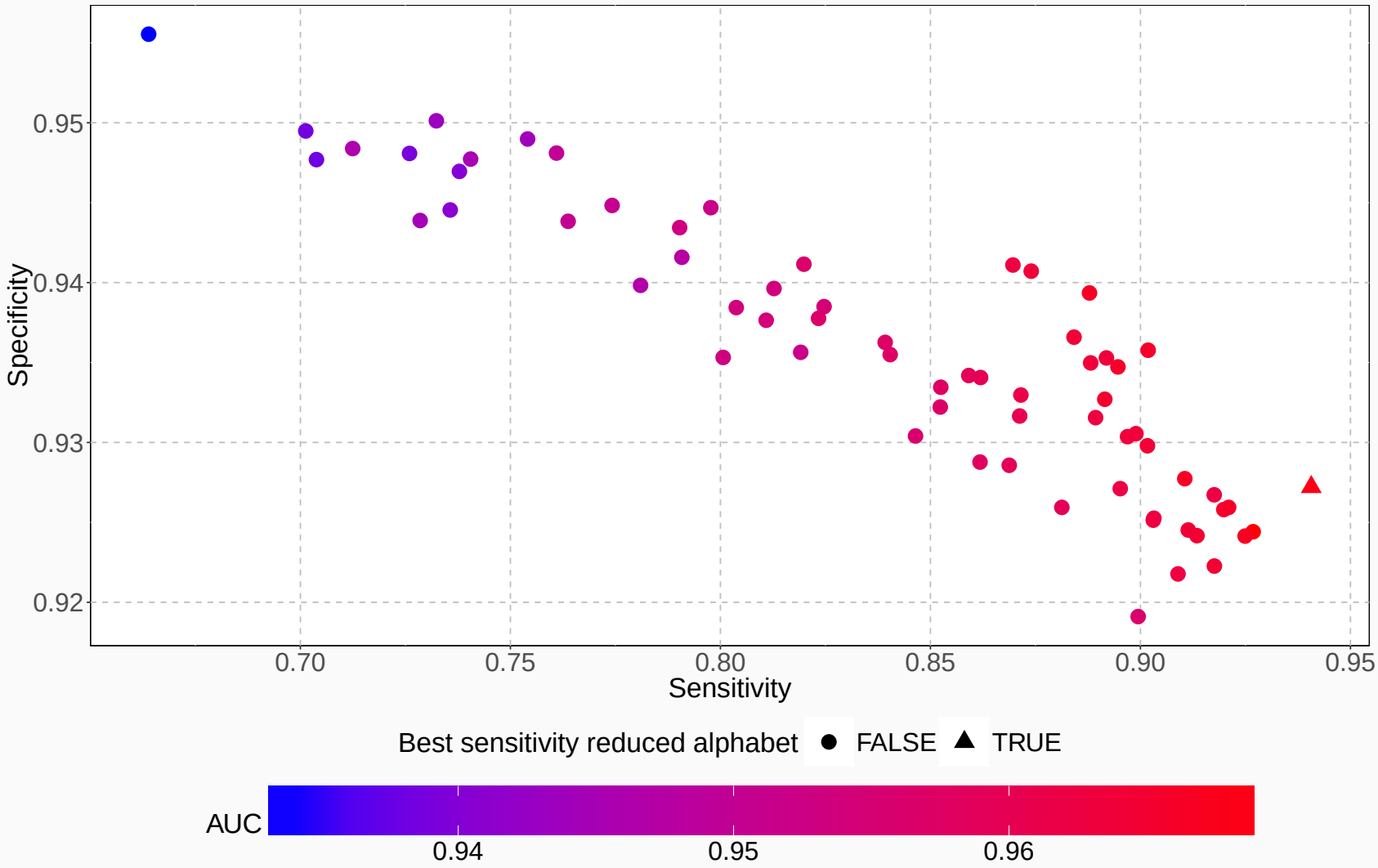
Cross-validation

We built 96 reduced amino acid alphabets (each based on one scale per a given property category) of length 4 (four distinct regions: n-, h-, c-region, mature protein).

Alphabets were evaluated in a cross-validation experiment using hidden semi-Markov models trained on eukaryotic sequences.

Results

Cross-validation



Best sensitivity encoding

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

Best sensitivity encoding

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

I. Charged or uncharged but polar amino acids absent in h-region.

Best sensitivity encoding

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

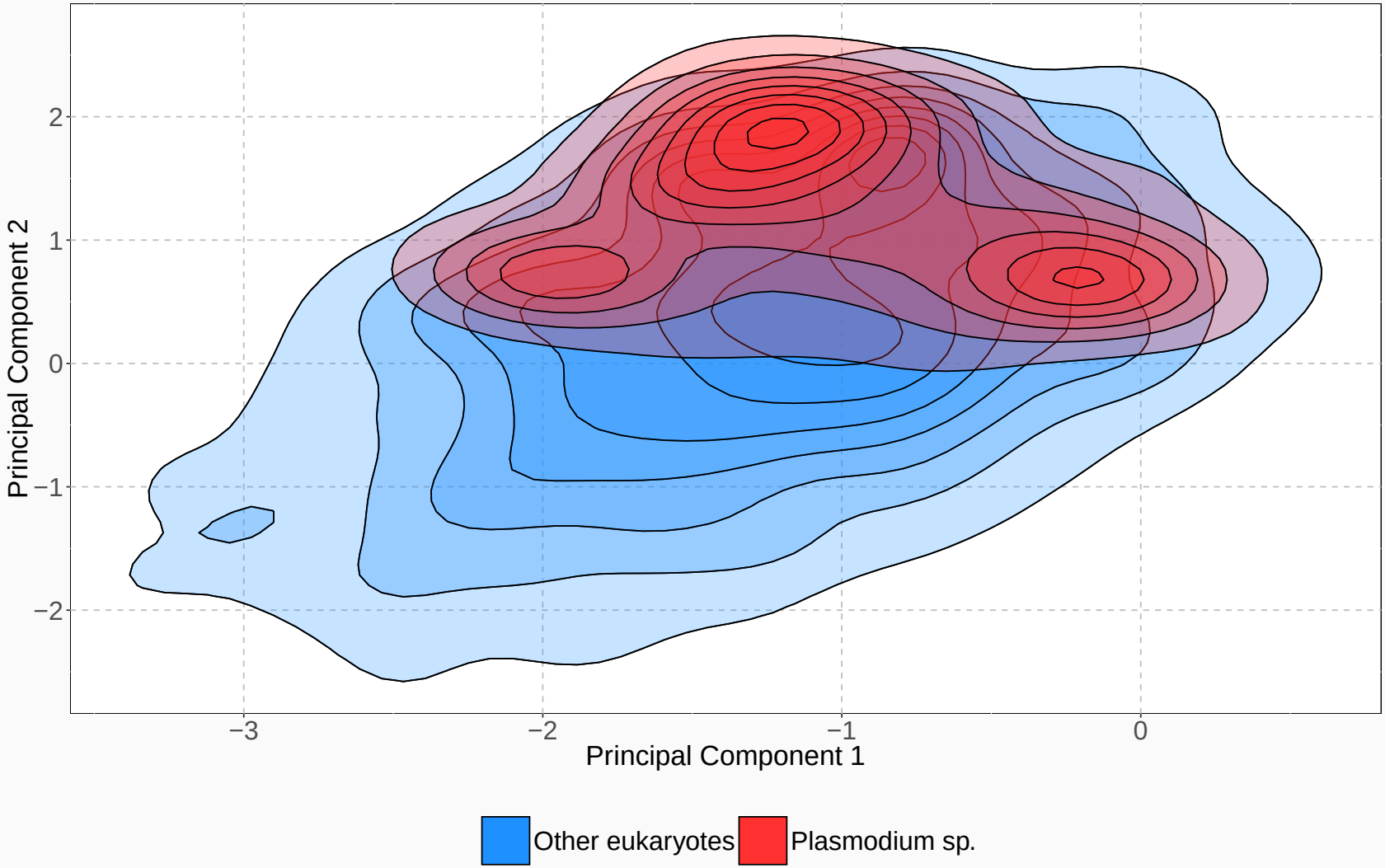
II. Polar and uncharged amino acids common in c-region.

Best sensitivity encoding

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

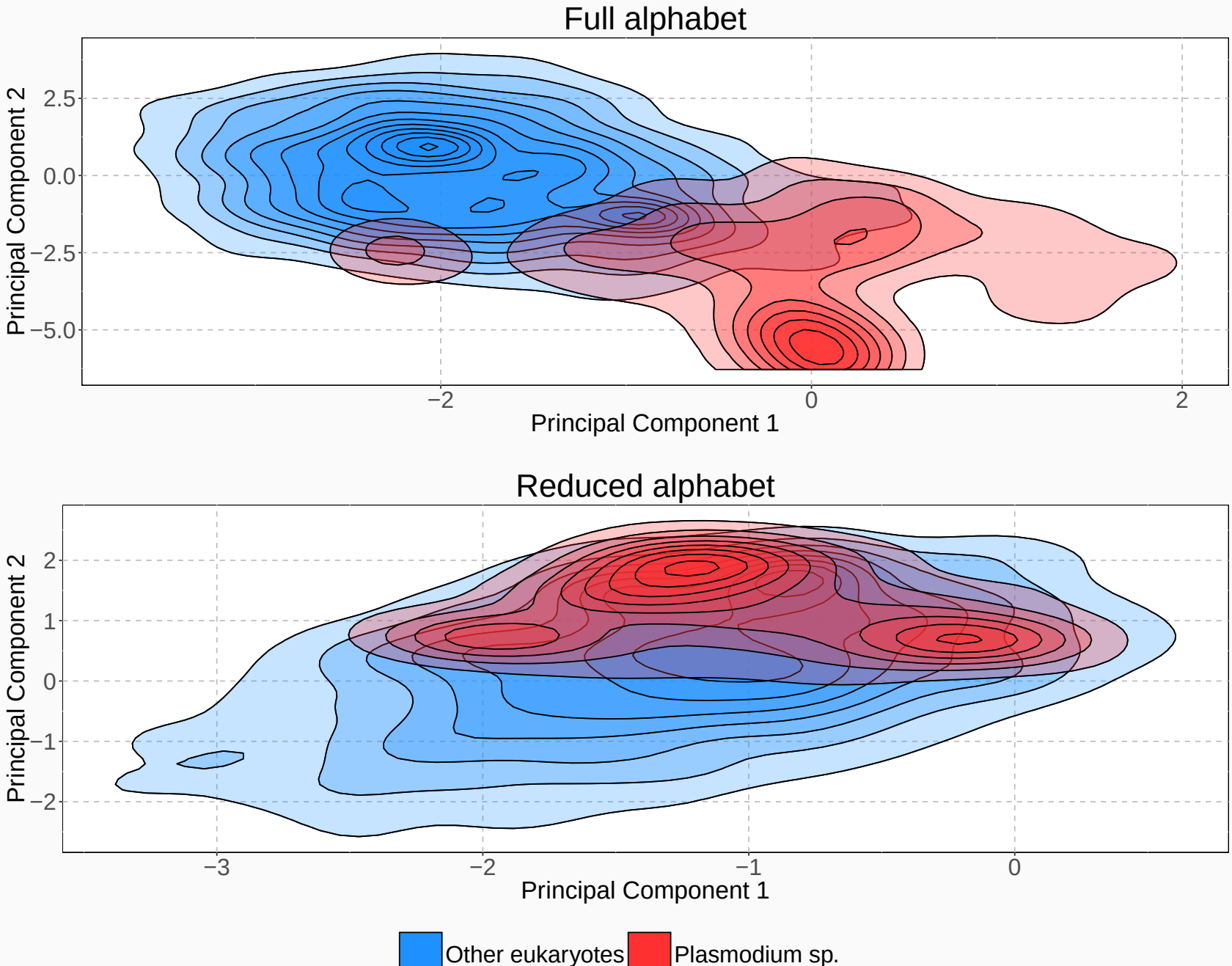
III. Hydrophobic amino acids common in h-region.

PCA of amino acid frequency



Signal peptides, after the reduction of the amino acid alphabet, group together despite their different origins.

PCA of amino acid frequency



Benchmark

Benchmark data set: 51 proteins with signal peptide and 211 proteins without signal peptide from members of *Plasmodiidae*.

Predictor: *signalHsmm*-2010, hidden semi-Markov model trained on data set of 3676 eukaryotic proteins with signal peptides added before year 2010 encoded using the best sensitivity reduced alphabet.

Benchmark

	MCC	AUC
signalP 4.1 (no tm) (Petersen et al., 2011)	0.6872	0.8667
signalP 4.1 (tm) (Petersen et al., 2011)	0.6196	0.7951
signalP 3.0 (NN) (Bendtsen et al., 2004)	0.7220	0.8938
signalP 3.0 (HMM) (Bendtsen et al., 2004)	0.5553	0.7734
Phobius (Käll et al., 2004)	0.5895	0.7880
signalHsmm-2010	0.7409	0.9262
signalHsmm-2010 (hom. 50%)	**0.7621**	**0.9384**
signalHsmm-2010 (full alphabet)	0.6853	0.8718

Full alphabet: no amino alphabet reduction. hom. 50%: 50% homology reduction in the learning data set.

Conclusions and perspectives

Eukaryotic signal peptides have very similar amino acid composition in their regions considering only the physicochemical properties of residues.

signalHsmm allows sensitive scanning of proteomes for potential drug targets whenever the protein of the interest is guided to specific subcellular compartments of pathogenic organisms.

Availability

signalHsmm web-server

`http://smorfland.uni.wroc.pl/shiny/signalHsmm`.

signalHsmm R package

`https://CRAN.R-project.org/package=signalHsmm`.

Acknowledgements and funding

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

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

References

Bendtsen, J. D., Nielsen, H., von Heijne, G., and Brunak, S. (2004). Improved prediction of signal peptides: Signalp 3.0. *Journal of Molecular Biology*, 340(4):783 – 795.

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

References II

Jain, R. G., Rusch, S. L., and Kendall, D. A. (1994). Signal peptide cleavage regions. functional limits on length and topological implications. *The Journal of Biological Chemistry*, 269(23):16305–16310.

Kalanon, M. and McFadden, G. I. (2010). Malaria, Plasmodium falciparum and its apicoplast. *Biochemical Society Transactions*, 38(3):775–782.

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.

References III

Moeller, L., Gan, Q., and Wang, K. (2009). A bacterial signal peptide is functional in plants and directs proteins to the secretory pathway. *Journal of Experimental Botany*, 60(12):3337–3352.

Nagano, R. and Masuda, K. (2014). Establishment of a signal peptide with cross-species compatibility for functional antibody expression in both escherichia coli and chinese hamster ovary cells. *Biochemical and Biophysical Research Communications*, 447(4):655 – 659.

References IV

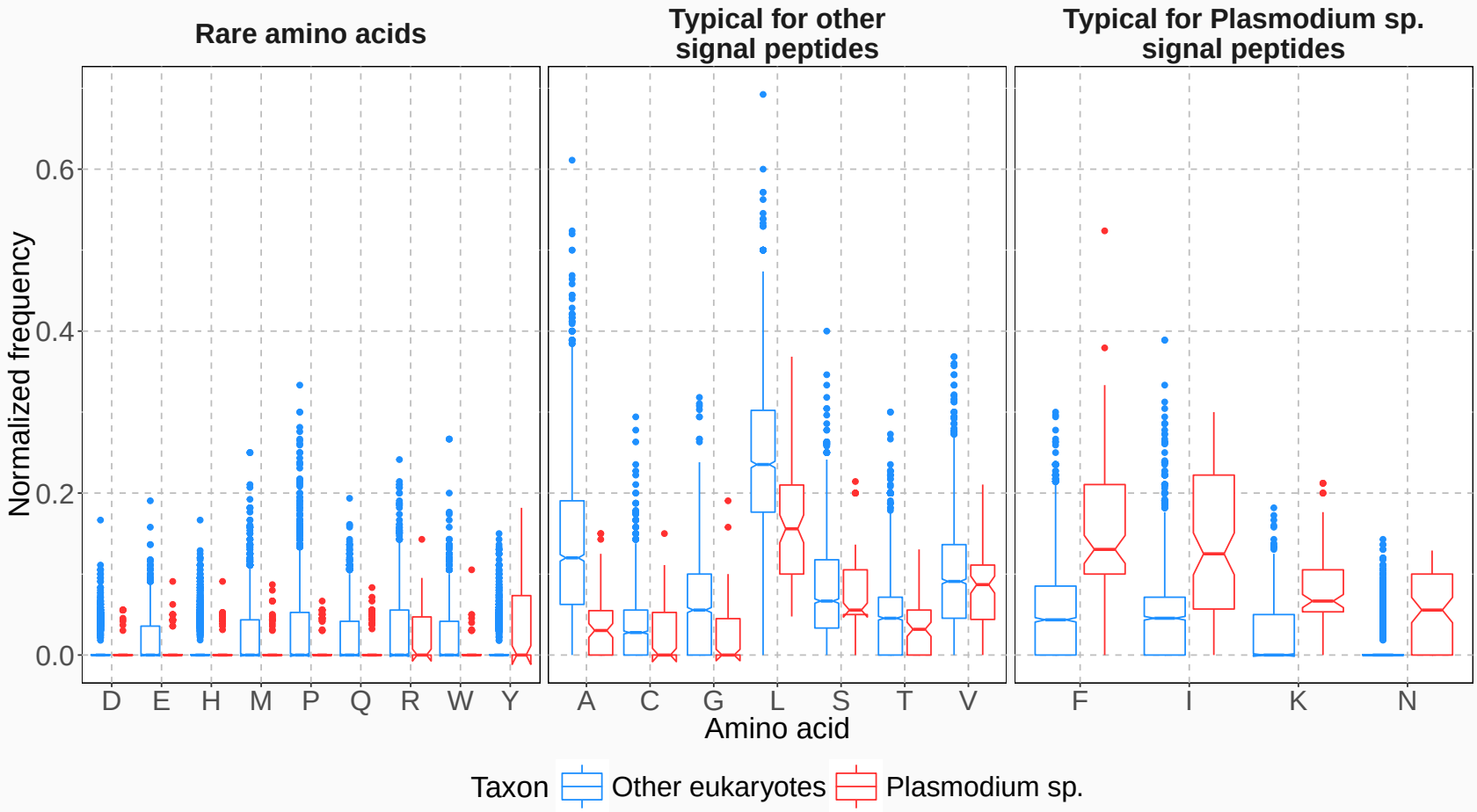
Nielsen, H. and Krogh, A. (1998). Prediction of signal peptides and signal anchors by a hidden markov model. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 6:122–130.

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.

Ralph, S. A., van Dooren, G. G., Waller, R. F., Crawford, M. J., Fraunholz, M. J., Foth, B. J., Tonkin, C. J., Roos, D. S., and McFadden, G. I. (2004). Tropical infectious diseases: Metabolic maps and functions of the Plasmodium falciparum apicoplast. *Nature Reviews Microbiology*, 2(3):203–216.

References V

Tonkin, C. J., Foth, B. J., Ralph, S. A., Struck, N., Cowman, A. F., and McFadden, G. I. (2008). Evolution of malaria parasite plastid targeting sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 105(12):4781–4785.



If notches are overlapping, two groups can be considered equal.