Prediction of malarial signal peptides using signalHsmm

Michał Burdukiewicz

October 5, 2016

1 Introduction

Heavy adenine-thymine bias of malarial genomes alters amino acid composition of malarial proteins, including signal peptides. Simple PCA analysis of amino acid frequency shows that signal peptides of *Plasmodiidae* do not group with signal peptides of other eukaryotes (Figure 1A).

The function of signal peptides enforces presence of amino acids with well defined physicochemical properties. Therefore, the difference between signal peptides of *Plasmodiidae* and other eukaryotes should disappear when we use reduced amino acid alphabet aggregating together residues with similar properties.

2 Reduced alphabet

We generated 96 reduced amino acid alphabets using combination of physicochemical properties relevant to signal peptide architecture (charge, polarity, hydrophobicity). To assess if reduced amino acid alphabets create more general model of signal peptides, we build a signal peptide predictor (based on hidden semi-Markov models) separately for each alphabet. In a cross-validation experiment (using only eukaryotic proteins) we find a reduced amino acid alphabet providing the best sensitivity (and second best AUC) (Table 1).

It is important to note, that reduced amino acid alphabet groups together signal peptides belonging to *Plasmodiidae* and other eukaryotes (Figure 1A).

Hidden semi-Markov model is a variation of Markov model used for example in signalP 2.0 and 3.0. We strayed from the pure Markov model, when we find that distributions of regional length is not exponential as implied by the Markov framework. Hidden semi-Markov algorithm allowed us to more precisely model regional structure of signal peptide offering an advantage over HMM implementations.

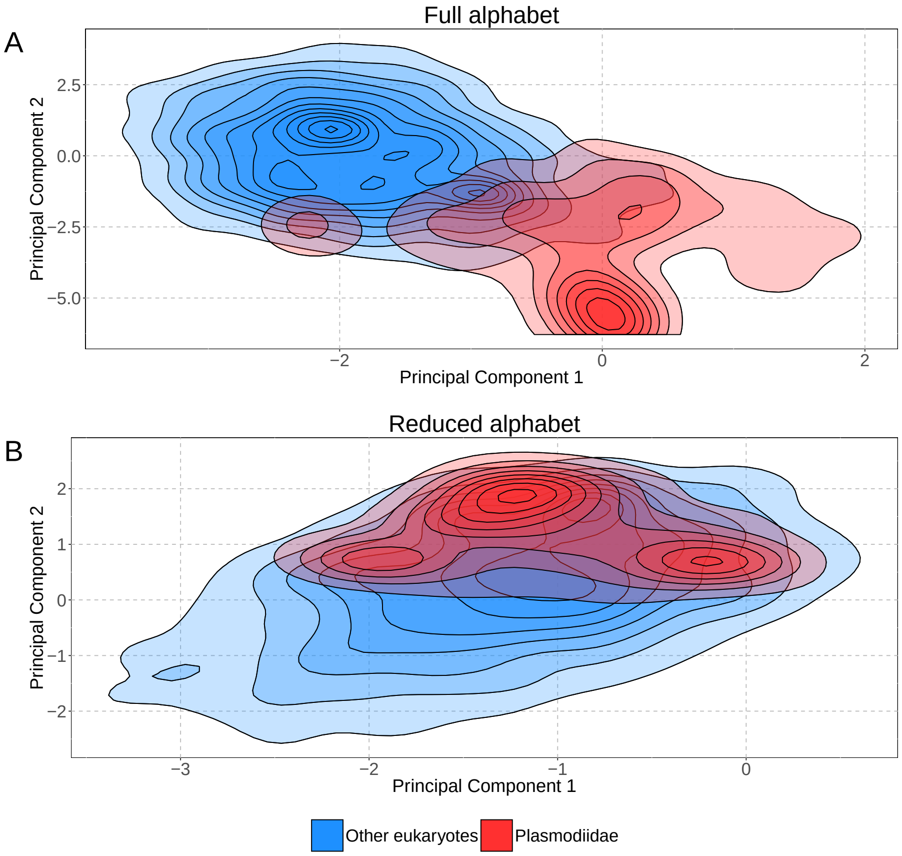

Figure 1: Principal component analysis of amino acid frequency in signal peptides belonging to *Plasmodiidae* and other eukaryotes. A) Frequency of amino acids. B) Frequency of amino acids encoded using the reduced alphabet.

Table 1: The best performing reduced amino acid 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

3 Benchmark

To create benchmark data set, we extracted proteins with signal peptide belonging to members of *Plasmodiidae* (51 proteins after 50% homology reduction). As negative data set we used proteins without signal peptide from the same taxon (211 proteins after 50% homology reduction).

As predictor we used *signalHsmm*-2010, hidden semi-Markov model trained on data set of 3,676 eukaryotic proteins with signal peptides added before year 2010 and encoded using the best sensitivity reduced alphabet. signalHsmm was compared to other predictors of signal peptides. As a negative control, we also benchmarked an iteration of signalHsmm that does not employ reduced amino acid alphabet.

Table 2: Results of benchmark. Full alphabet: no amino alphabet reduction.

	Sensitivity	Specificity	MCC	AUC
signalP 4.1 (no tm) (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
signalP 3.0 (NN) (Bendtsen et al., 2004)	0.8824	0.9052	0.7220	0.8938
signalP 3.0 (HMM) (Bendtsen et al., 2004)	0.6275	0.9194	0.5553	0.7734
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-2010	**0.9804**	0.8720	**0.7409**	**0.9262**
signalHsmm-2010 (full alphabet)	0.8431	0.9005	0.6853	0.8718

4 Workplan

Aim: improve performance of signalP for atypical signal peptides using a reduced amino acid alphabet.

Since signalP 5 is not yet available, I want to instead use signalP 4.1.

1. Adjust signalP 4.1 for reduced alphabets. I don't have an access to signalP 4.1 source code, but it is possible, that it has hardcoded alphabet of 20 amino acids, so a bit of programming might be necessary.

2. Benchmark modified signalP 4.1 on external data set(s) of atypical signal peptides and compare with normal signalP 4.1 and signalP 3.0.

If modified signalP 4.1 is as accurate as it was during benchmark on proteins belonging to other eukaryotes and it predicts atypical signal peptides better, it gives us a rationale to use reduced alphabets and justifies further search for more optimized reduced amino acid alphabet.

5 Concerns

It is possible that reduction of the alphabet may improve the general performance of signal peptide prediction, but lower accuracy of cleavage sites prediction. Cleavage sites seem to require more defined motifs than whole signal peptide and may need larger alphabets, possibly even the full amino acid alphabet.

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.

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.