

# Expanding signalHsmm using n-grams

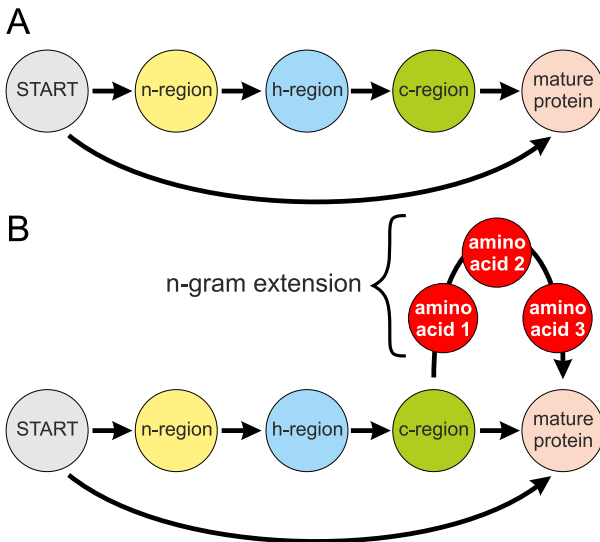
Michał Burdukiewicz, Piotr Sobczyk

University of Wrocław, Department of Genomics, Poland

# Outline

- 1 Motivation
- 2 Impact of the n-gram extension
- 3 Differences in cleavage site composition
  - Amino acid sequences
  - Degenerated amino acid sequences
- 4 Heuristic algorithm
- 5 Summary table

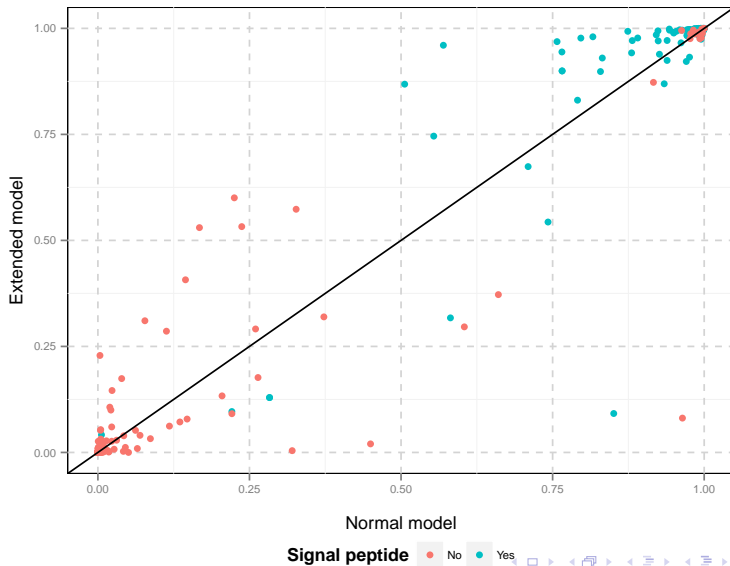
To improve our model, we added few supplementary states representing specific motives that might occur in the proximity of cleavage site. The structure of cleavage sites, more conserved than other parts of signal peptide, may be reflected by n-grams (k-mers), short vectors of  $n$  characters derived from input sequences.



The diagram of simple (A) and extended version of signalHsmm with the

# Outline

- 1 Motivation
- 2 Impact of the n-gram extension
- 3 Differences in cleavage site composition
  - Amino acid sequences
  - Degenerated amino acid sequences
- 4 Heuristic algorithm
- 5 Summary table



Performance of different classifiers.

	AUC	TP	TN	FP	FN
signalPnotm	0.9416	208	195	19	6
signalPtm	0.9673	205	209	5	9
predsi	0.8949	194	189	25	20
phobius	0.9509	207	200	14	7
philius	0.9369	204	197	17	10
signalHsmm2010	0.9526	198	191	23	16
signalHsmm1989	0.9562	202	194	20	12
signalKmer	0.9695	206	194	20	8

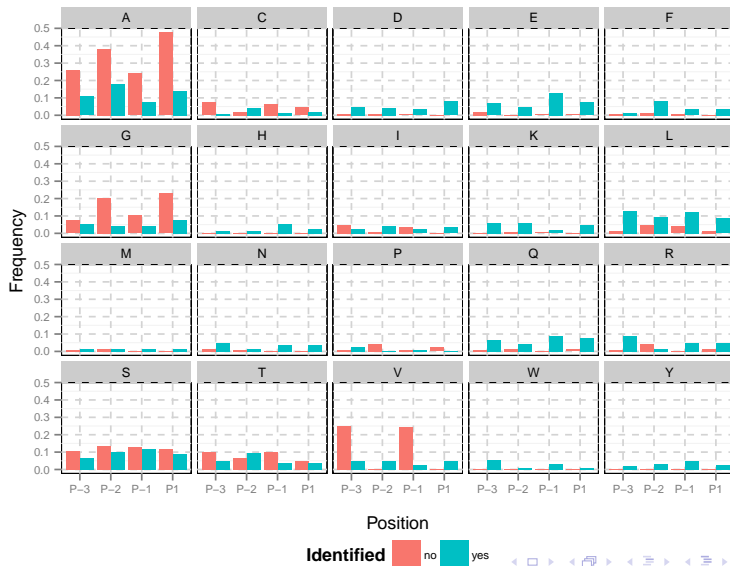
# Outline

- 1 Motivation
- 2 Impact of the n-gram extension
- 3 Differences in cleavage site composition
  - Amino acid sequences
  - Degenerated amino acid sequences
- 4 Heuristic algorithm
- 5 Summary table

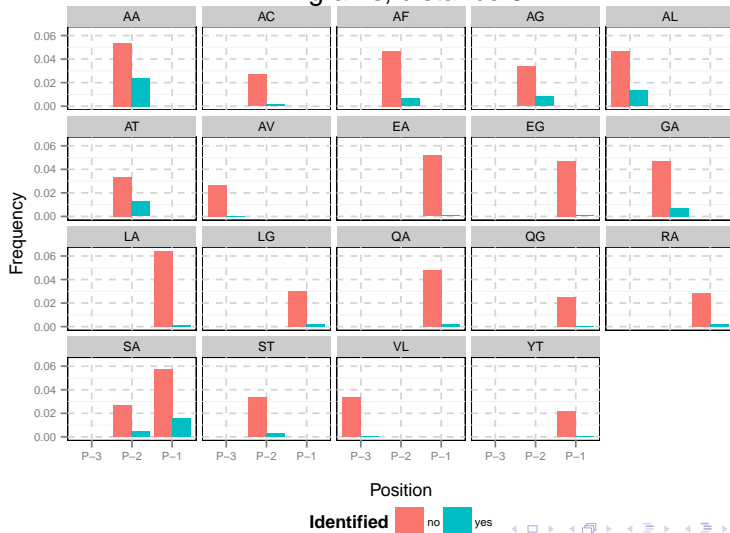


- 1 Train signalHsmm on all sequences with signal peptide.
- 2 Test signalHsmm on training set.
- 3 Assign label "not recognized" to not recognized sequences.

Considering 2525 sequences, 149 (5.9%) are wrongly recognized.



## 2-grams, distance 0



## 2-grams, distance 1



---

### Groups

---

D, E, H, K, N, Q, R

G, P, S, T, Y

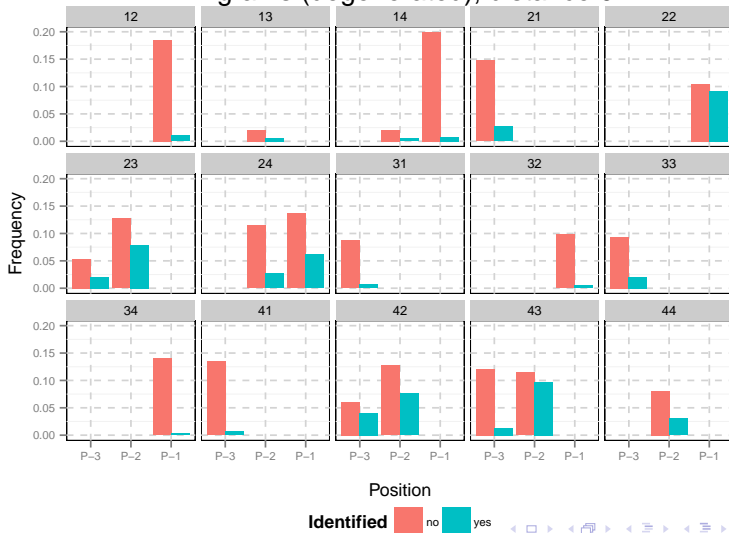
F, I, L, M, V, W

A, C

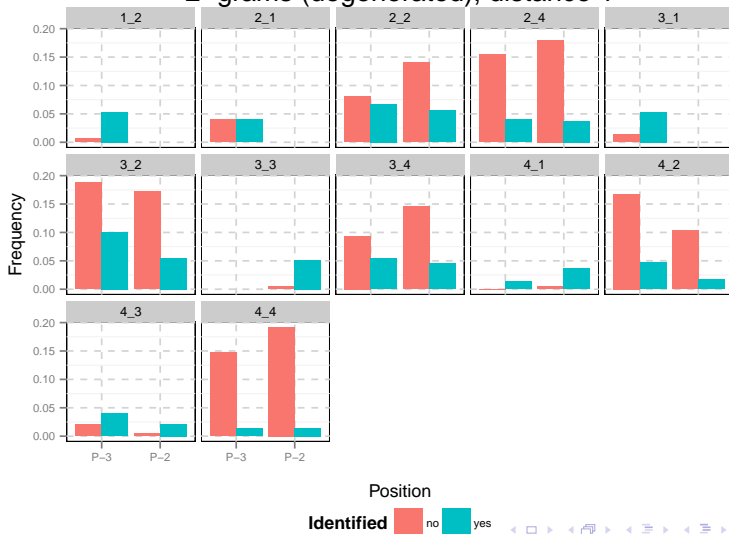
---

Classification of amino acids used by signalHsmm.

## 2-grams (degenerated), distance 0



## 2-grams (degenerated), distance 1



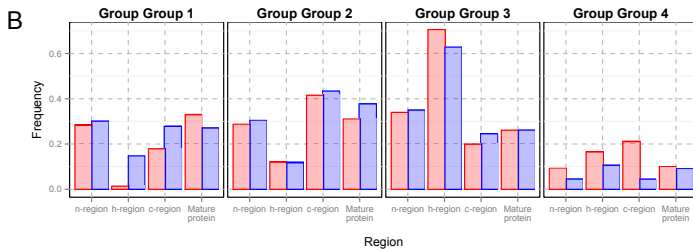
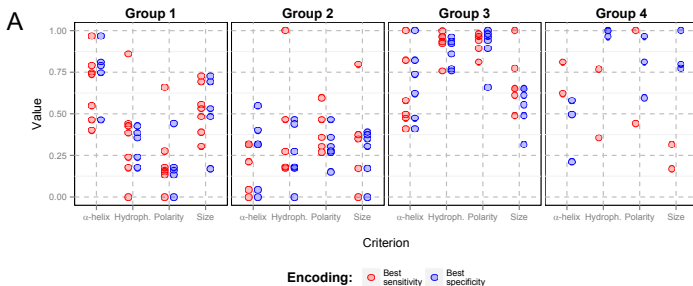


# Outline

- 1 Motivation
- 2 Impact of the n-gram extension
- 3 Differences in cleavage site composition
  - Amino acid sequences
  - Degenerated amino acid sequences
- 4 Heuristic algorithm**
- 5 Summary table

The heuristic algorithm relies on presence of some amino acids to define borders of regions.

According to the original authors, C-region is recognized by a presence of: R, H, K, D, E.



---

### Groups

---

D, E, H, K, N, Q, R

G, P, S, T, Y

F, I, L, M, V, W

A, C

---

Classification of amino acids used by signalHsmm.

After addition of G, S, T, P, Y, 135 (5.3%) sequences are wrongly recognized (compared to 149).

# Outline

- 1 Motivation
- 2 Impact of the n-gram extension
- 3 Differences in cleavage site composition
  - Amino acid sequences
  - Degenerated amino acid sequences
- 4 Heuristic algorithm
- 5 Summary table

## Performance of different classifiers.

	AUC	TP	TN	FP	FN
signalPnotm	0.9416	208	195	19	6
signalPtm	0.9673	205	209	5	9
predsi	0.8949	194	189	25	20
phobius	0.9509	207	200	14	7
philius	0.9369	204	197	17	10
signalHsmm2010	0.9526	198	191	23	16
signalHsmm1989	0.9562	202	194	20	12
signalKmer	0.9695	206	194	20	8
signalHsmmH	0.9458	206	188	26	8
signalHsmmHKmer	0.9653	209	193	21	5