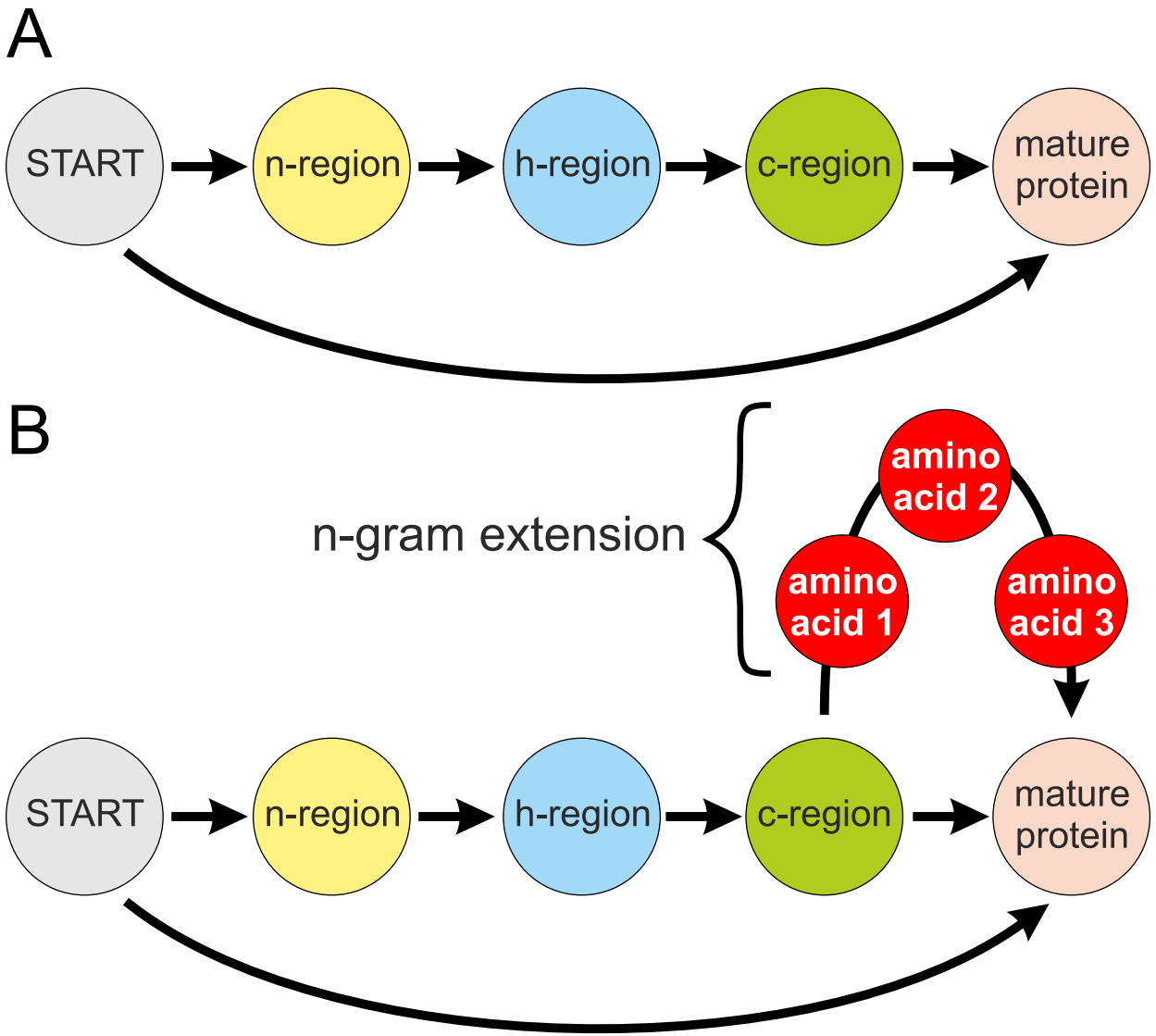
Expanding signalHsmm using n-grams

Michał Burdukiewicz, Piotr Sobczyk

University of Wrocław, Department of Genomics, Poland

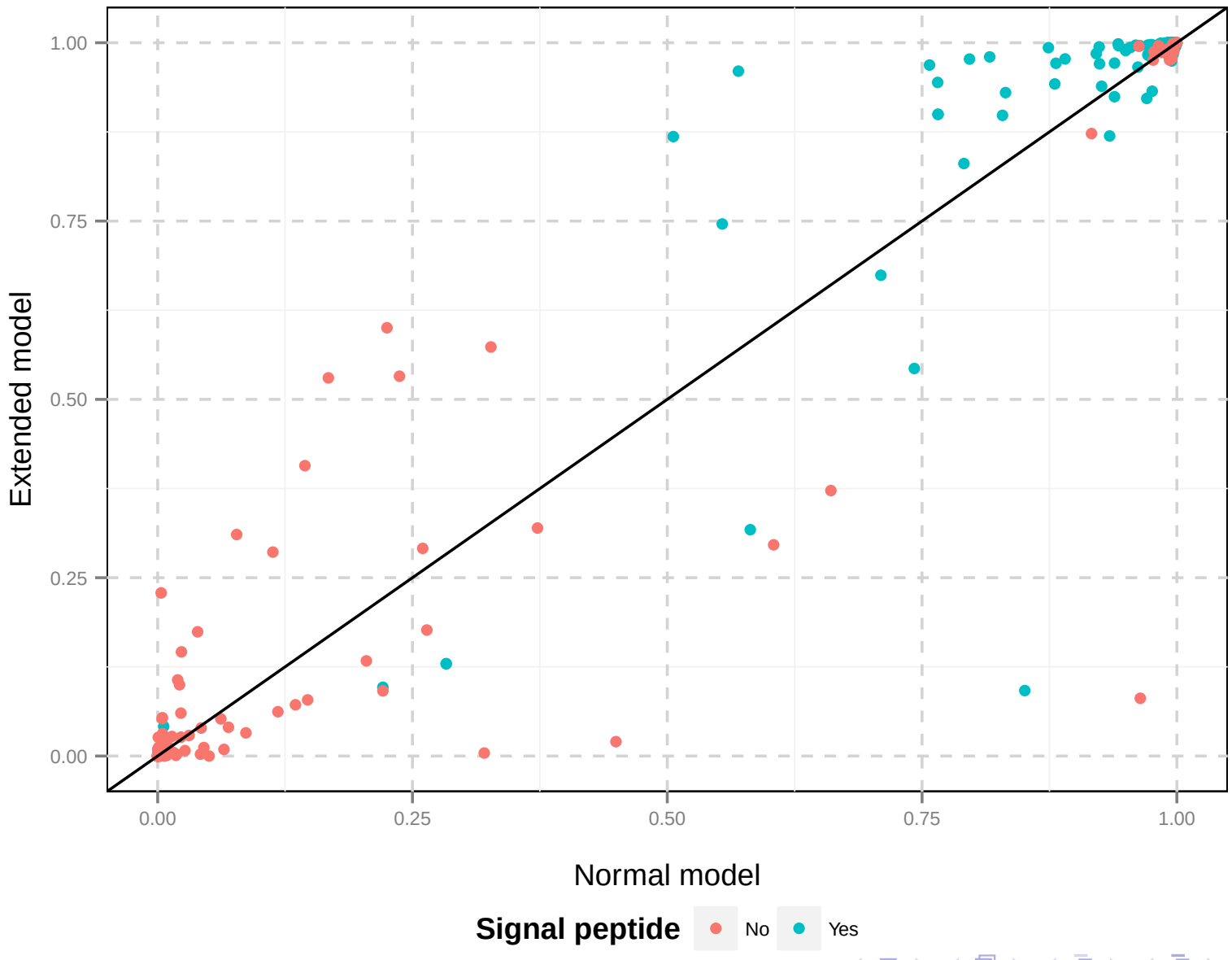
Outline

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 n-gram cleavage site model (B).

Outline


Extended model
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Normal model
Signal peptide
No
Yes

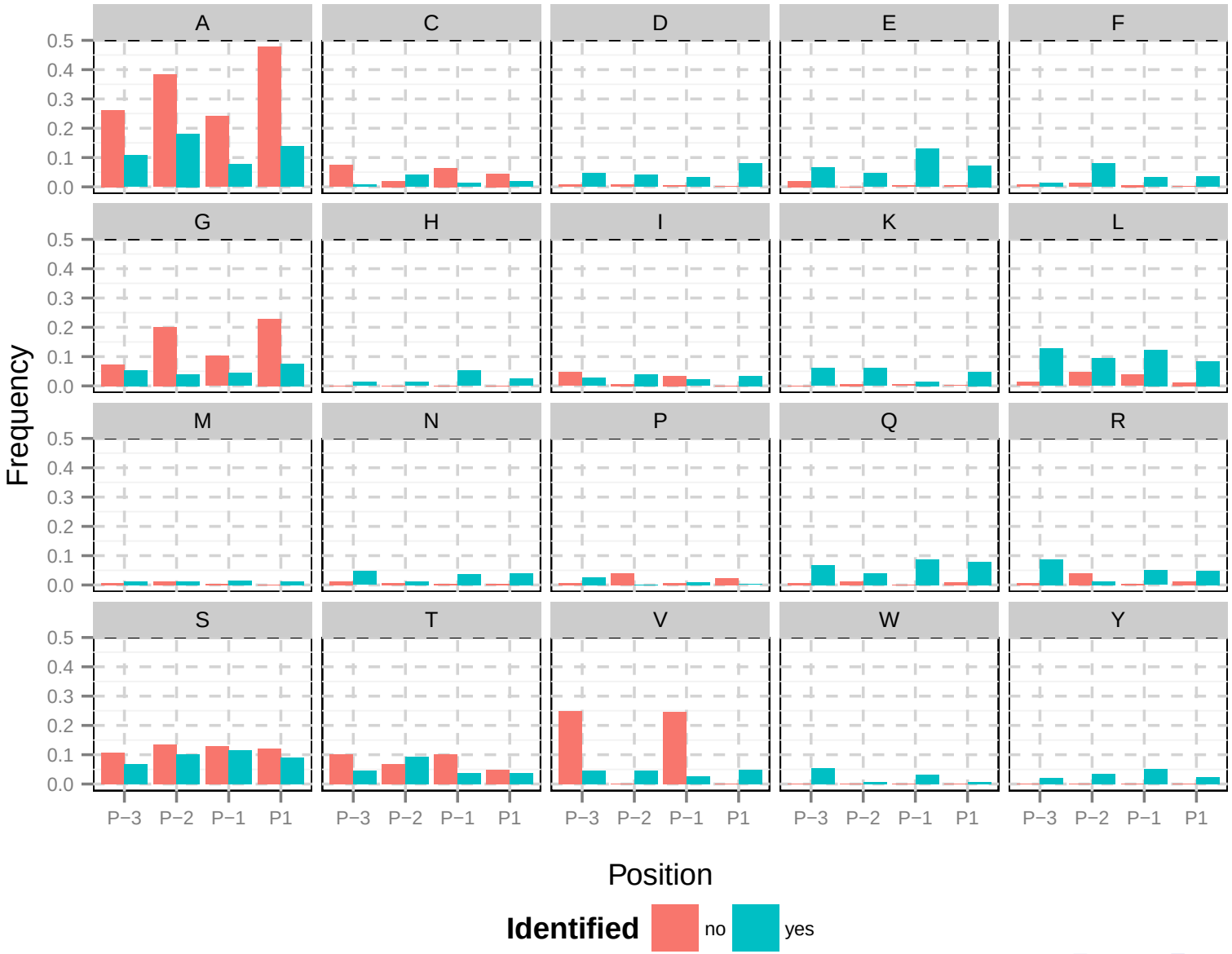
Performance of different classifiers.

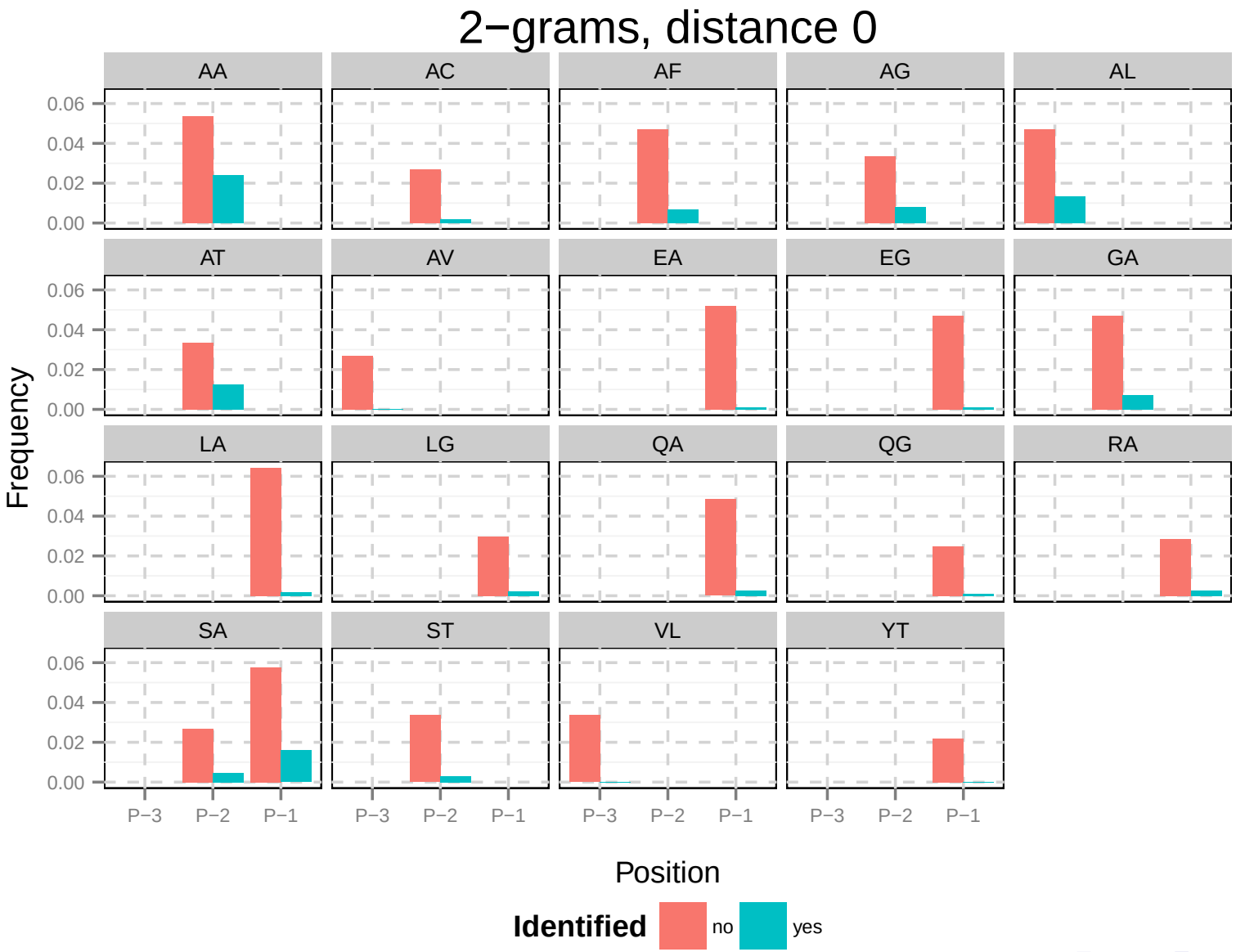
	AUC	TP	TN	FP	FN
signalPnotm	0.9416	208	195	19	6
signalPtm	0.9673	205	209	5	9
predsi	0.8949	194	189	25	20
phobius	0.9509	207	200	14	7
philius	0.9369	204	197	17	10
signalHsmm2010	0.9526	198	191	23	16
signalHsmm1989	0.9562	202	194	20	12
signalKmer	0.9695	206	194	20	8

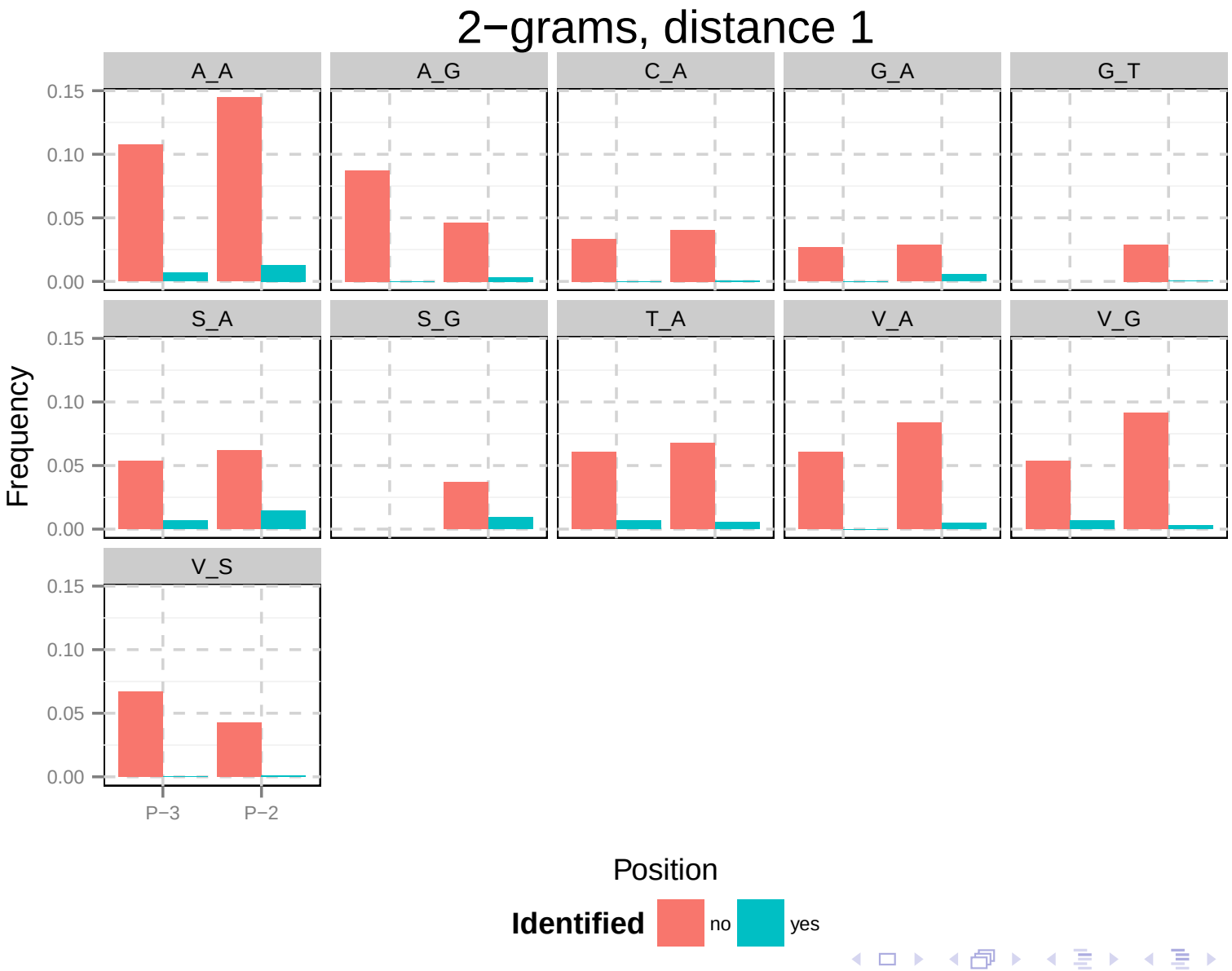
Outline

1. 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.

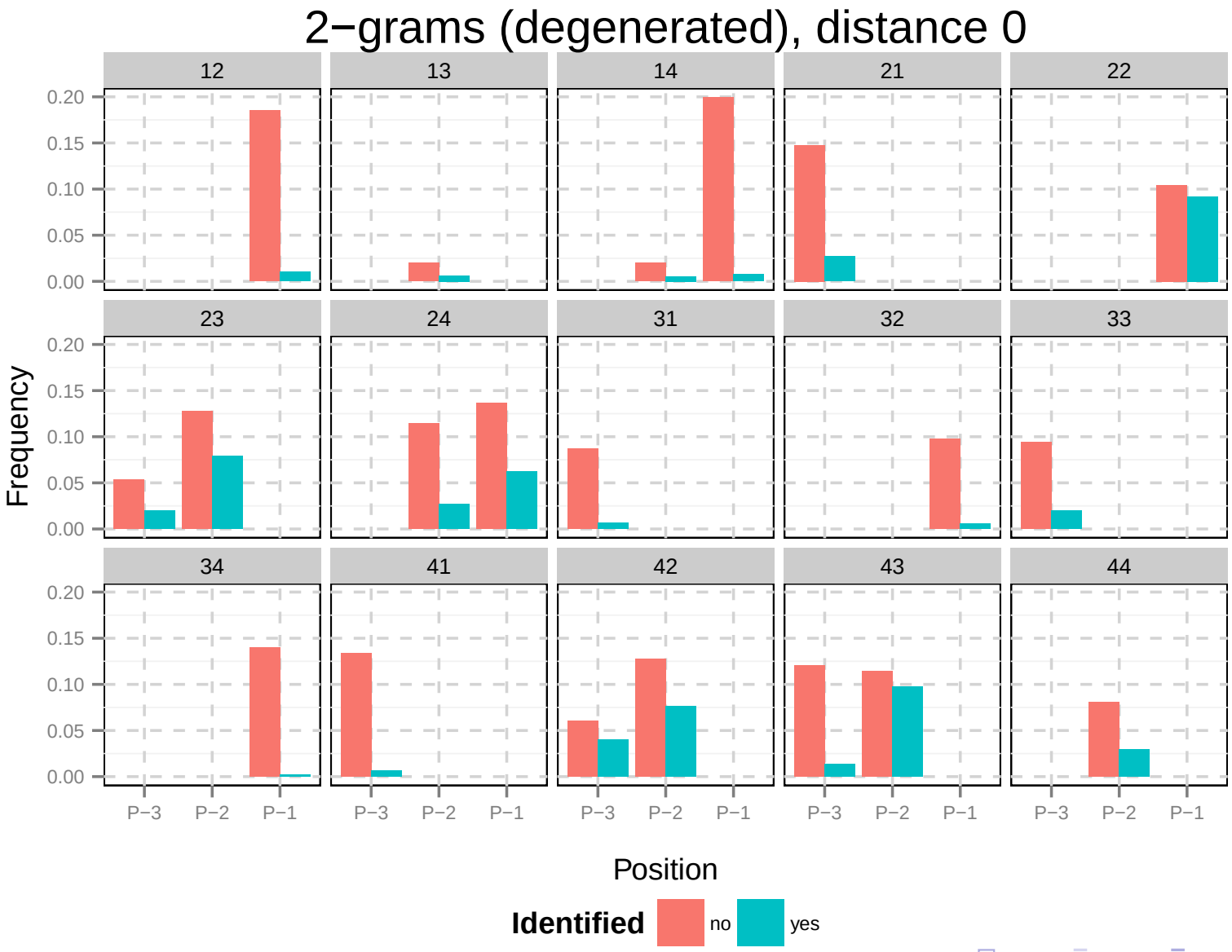

A
C
D
E
F
G
H
I
K
L
M
N
P
Q
R
S
T
V
W
Y
Frequency
0.0
0.1
0.2
0.3
0.4
0.5
P−3
P−2
P−1
P1
Position
Identified
no
yes

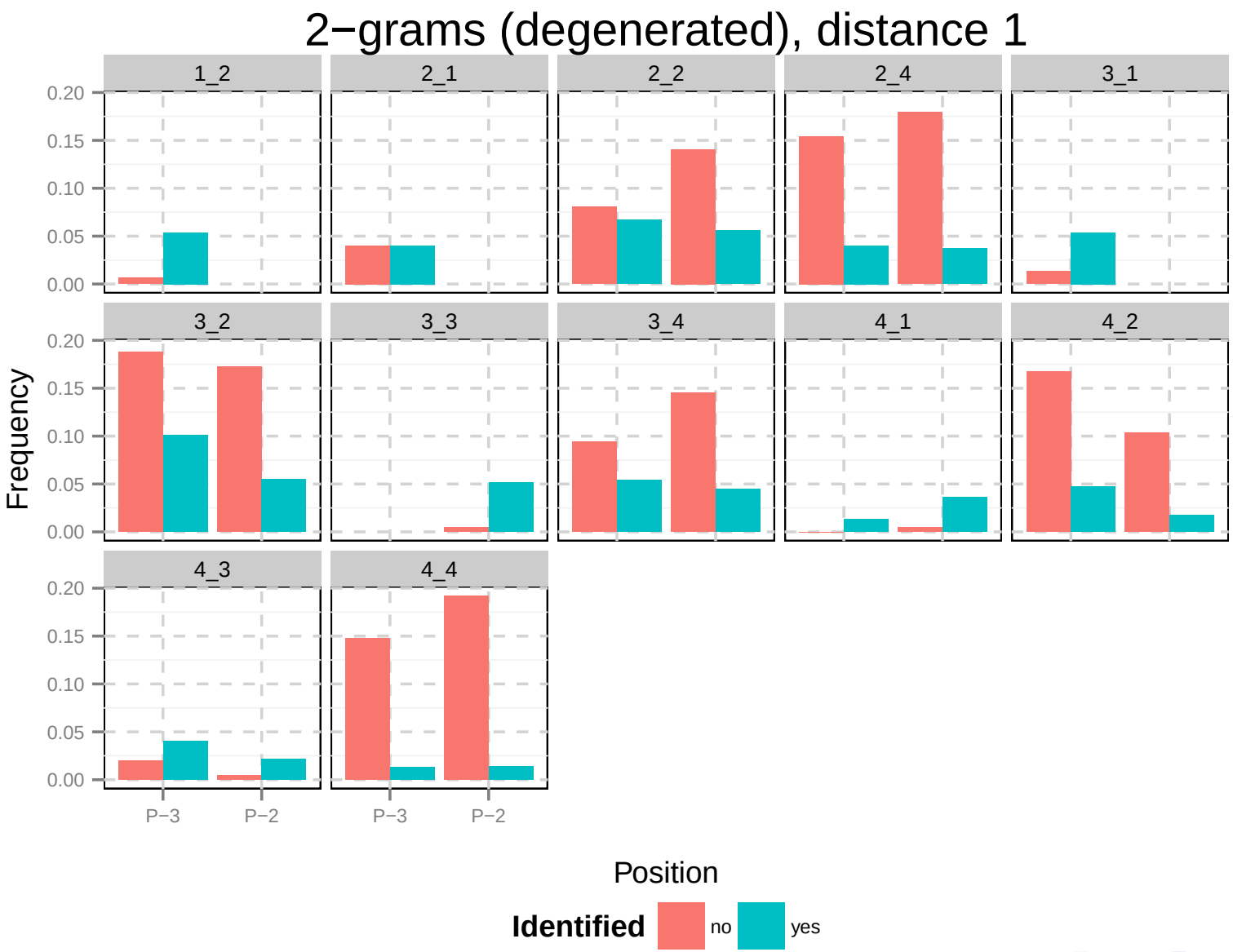

2−grams, distance 0
AA
AC
AF
AG
AL
AT
AV
EA
EG
GA
LA
LG
QA
QG
RA
SA
ST
VL
YT
0.06
0.04
0.02
0.00
Frequency
P−3
P−2
P−1
Position
Identified
no
yes


2−grams, distance 1
A_A
A_G
C_A
G_A
G_T
S_A
S_G
T_A
V_A
V_G
V_S
Frequency
0.15
0.10
0.05
0.00
P−3
P−2
Position
Identified
no
yes

Groups
D, E, H, K, N, Q, R
G, P, S, T, Y
F, I, L, M, V, W
A, C

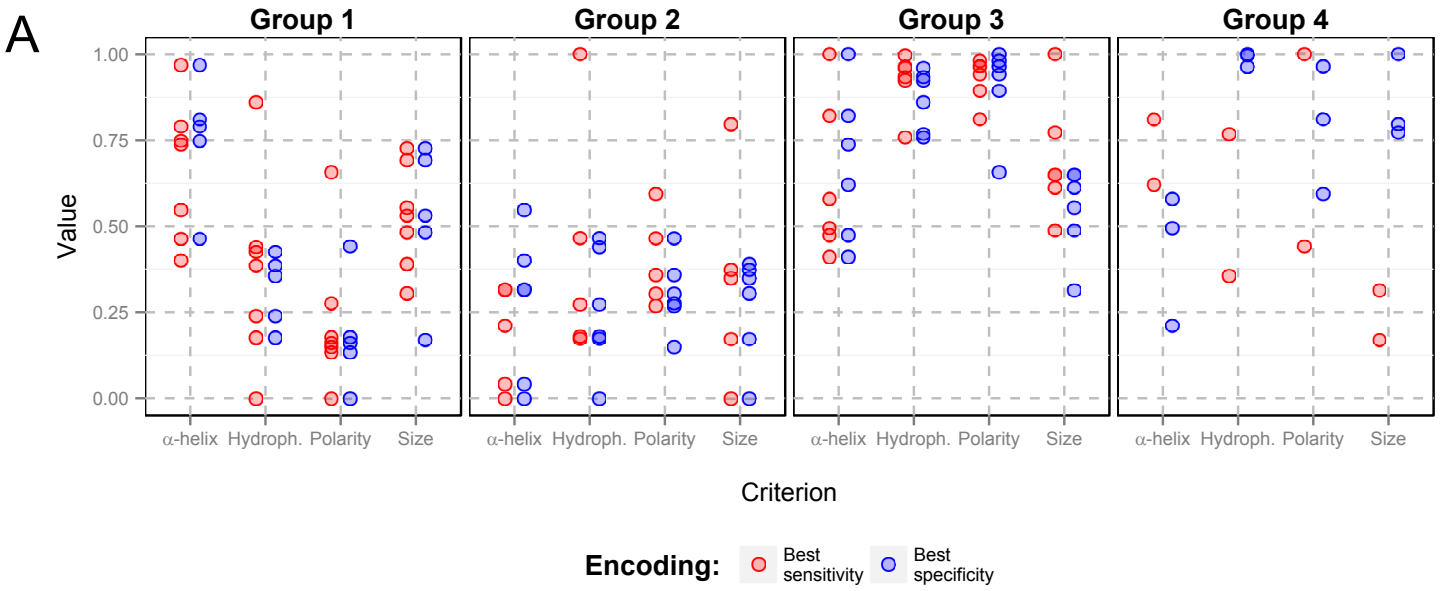
Classification of amino acids used by signalHsmm.

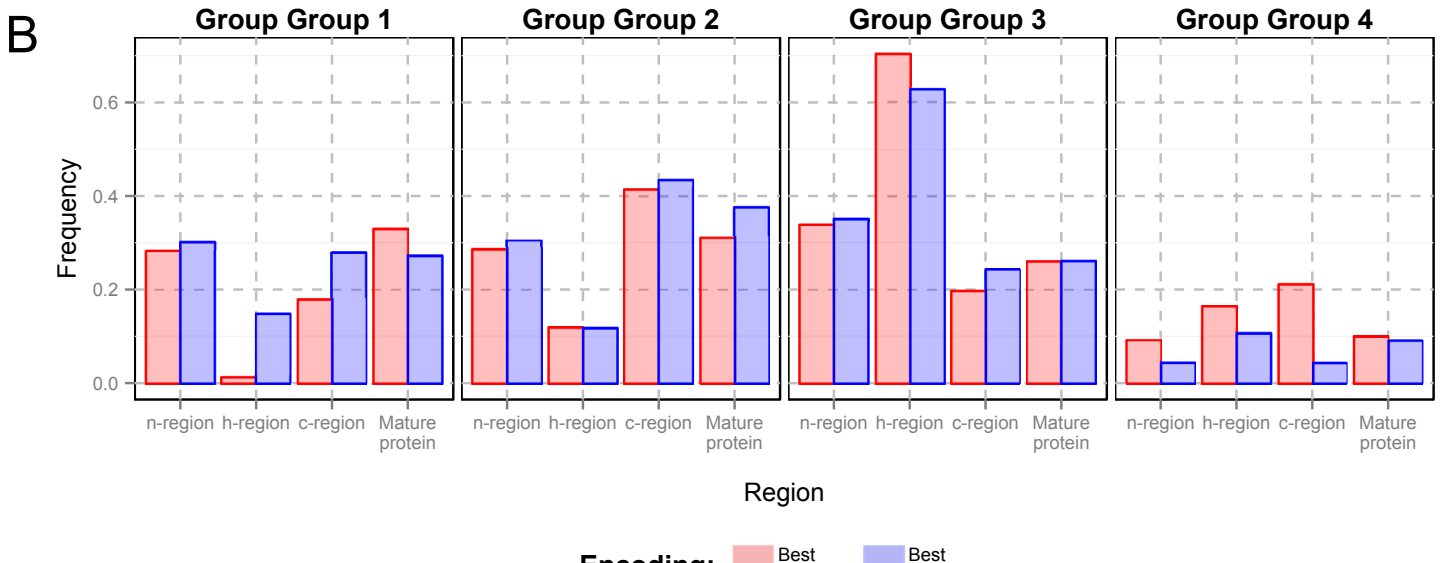

2−grams (degenerated), distance 0
12
13
14
21
22
23
24
31
32
33
34
41
42
43
44
Frequency
0.20
0.15
0.10
0.05
0.00
P−3
P−2
P−1
Position
Identified
no
yes


2−grams (degenerated), distance 1
1_2
2_1
2_2
2_4
3_1
3_2
3_3
3_4
4_1
4_2
4_3
4_4
Frequency
0.20
0.15
0.10
0.05
0.00
P−3
P−2
Position
Identified
no
yes

Outline

The heuristic algorithm relies on presence of some amino acids to define borders of regions.
According to the origingal authors, C-region is recognized by a presence of: R, H, K, D, E.


A
Group 1
Group 2
Group 3
Group 4
Value
1.00
0.75
0.50
0.25
0.00
α-helix
Hydroph.
Polarity
Size
Criterion
Encoding:
Best sensitivity
Best specificity


B
Group Group 1
Group Group 2
Group Group 3
Group Group 4
Frequency
0.6
0.4
0.2
0.0
n-region
h-region
c-region
Mature protein
Region
Encoding:
Best
Best

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).