Encodings of amino acids and their impact on signal peptide prediction

Michał Burdukiewicz

University of Wrocław, Department of Genomics, Poland

Outline

1. Motivation
2. Amino acid properties
    - Size
    - Polarity
    - PI
    - Hydrophobicity
    - α-chain
3. Methods
    - Amino acid groupings
    - Cross-validation
4. Cross-validation
    - AUC
    - H
    - Specificity
    - Mean cleavage site displacement
5. Conclusion and perspectives

To reduce dimensionality of the problem, signalHsmm aggregates amino acids to four physicochemical groups.

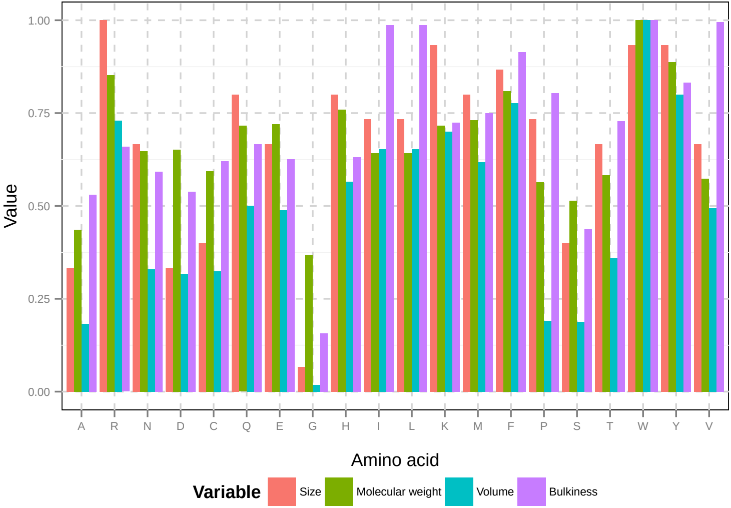
Groups
H, K, R
C, F, I, L, M, U, V, W
N, Q, S, T
A, D, E, G, P, Y

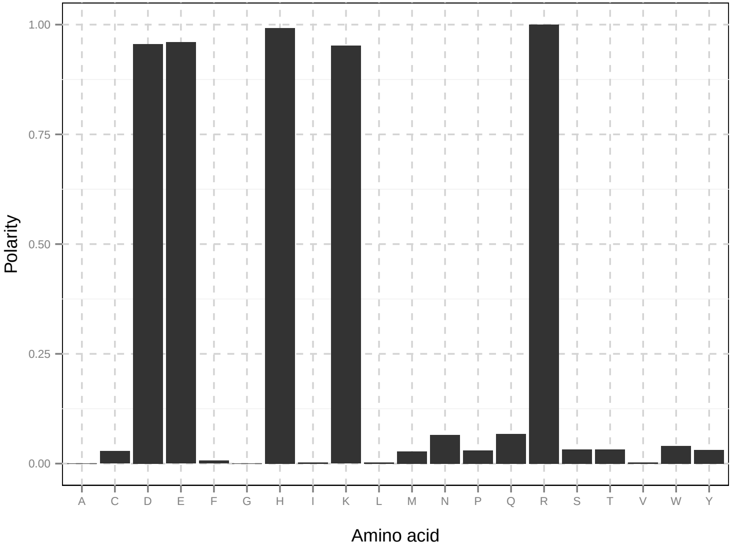
Classification of amino acids used by signalHsmm.

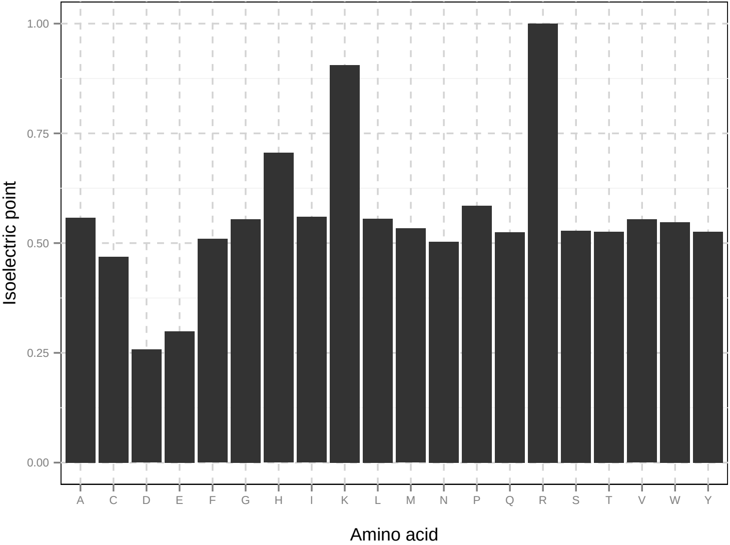
How optimal is this grouping? How does grouping of amino acids influence detection of signal peptides and cleavage sites?

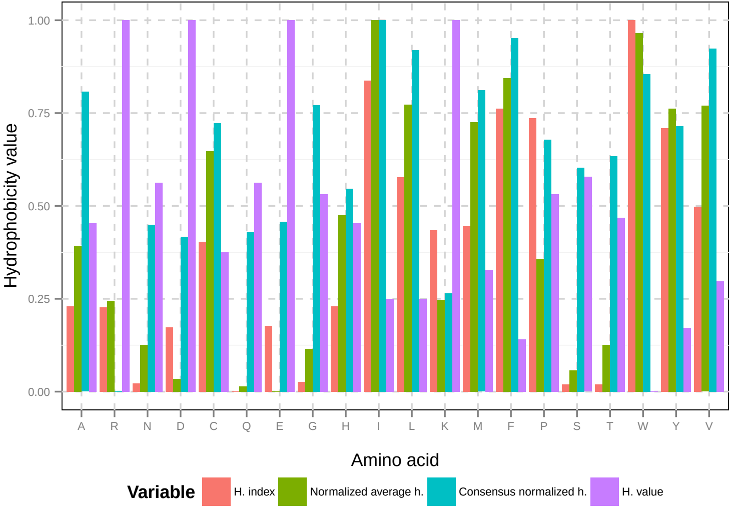
Outline

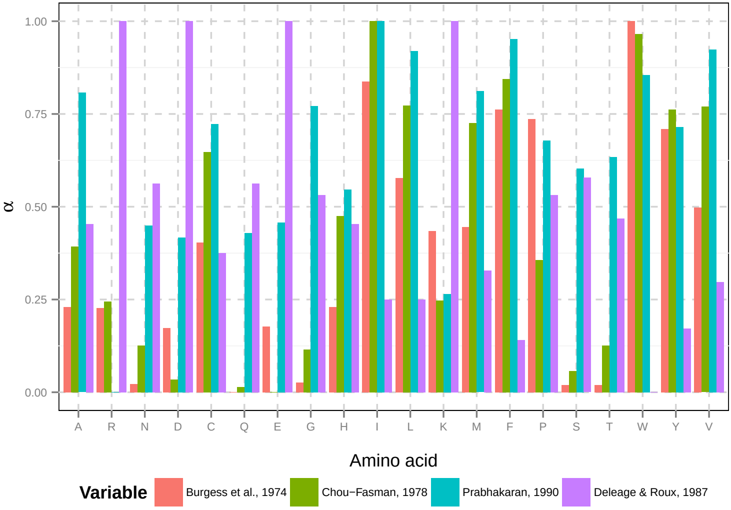
1. Motivation
2. Amino acid properties
   - Size
   - Polarity
   - PI
   - Hydrophobicity
   - α-chain
3. Methods
   - Amino acid groupings
   - Cross-validation
4. Cross-validation
   - AUC
   - H
   - Specificity
   - Mean cleavage site displacement
5. Conclusion and perspectives


Value
1.00
0.75
0.50
0.25
0.00
A
R
N
D
C
Q
E
G
H
I
L
K
M
F
P
S
T
W
Y
V
Amino acid
Variable
Size
Molecular weight
Volume
Bulkiness


Polarity
1.00
0.75
0.50
0.25
0.00
A
C
D
E
F
G
H
I
K
L
M
N
P
Q
R
S
T
V
W
Y
Amino acid


Isoelectric point
1.00
0.75
0.50
0.25
0.00
A
C
D
E
F
G
H
I
K
L
M
N
P
Q
R
S
T
V
W
Y
Amino acid


Hydrophobicity value
1.00
0.75
0.50
0.25
0.00
A
R
N
D
C
Q
E
G
H
I
L
K
M
F
P
S
T
W
Y
V
Amino acid
Variable
H. index
Normalized average h.
Consensus normalized h.
H. value


1.00
0.75
0.50
0.25
0.00
α
A
R
N
D
C
Q
E
G
H
I
L
K
M
F
P
S
T
W
Y
V
Amino acid
Variable
Burgess et al., 1974
Chou−Fasman, 1978
Prabhakaran, 1990
Deleage & Roux, 1987

Outline

1. Motivation
2. Amino acid properties
    - Size
    - Polarity
    - PI
    - Hydrophobicity
    - α-chain
3. Methods
    - Amino acid groupings
    - Cross-validation
4. Cross-validation
    - AUC
    - H
    - Specificity
    - Mean cleavage site displacement
5. Conclusion and perspectives

For each combination of each variant of all properties:

1. calculate euclidean distance between amino acids;
2. cluster amino acids (complete-linkage clustering);
3. extract four highest clusters of amino acids.

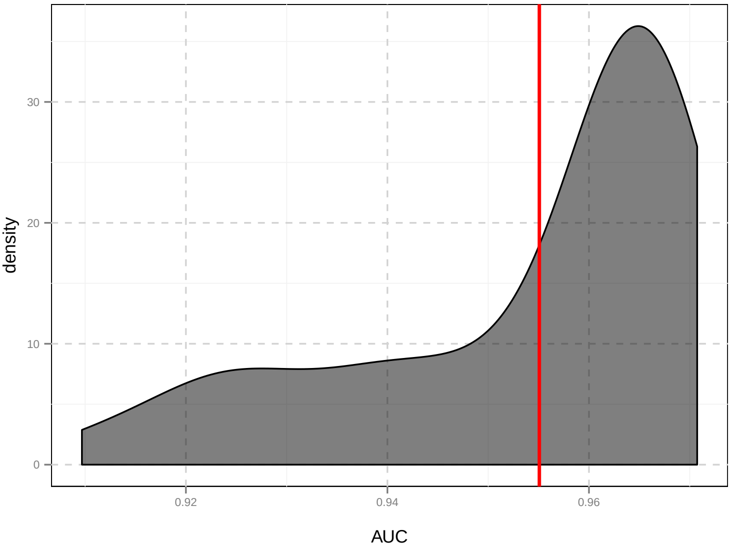
67 amino acid groupings in total (signalHsmm, 2 standard, 65 created).

For each grouping:

1. sample 3722 from 134044 negative sequences (to have balanced data set);
2. perform 5-fold cross-validation using all groupings (train an instance of signalHsmm using a single grouping);
3. repeat steps 1:2 15 times.

Outline

1. Motivation
2. Amino acid properties
   - Size
   - Polarity
   - PI
   - Hydrophobicity
   - α-chain
3. Methods
   - Amino acid groupings
   - Cross-validation
4. Cross-validation
   - AUC
   - H
   - Specificity
   - Mean cleavage site displacement
5. Conclusion and perspectives

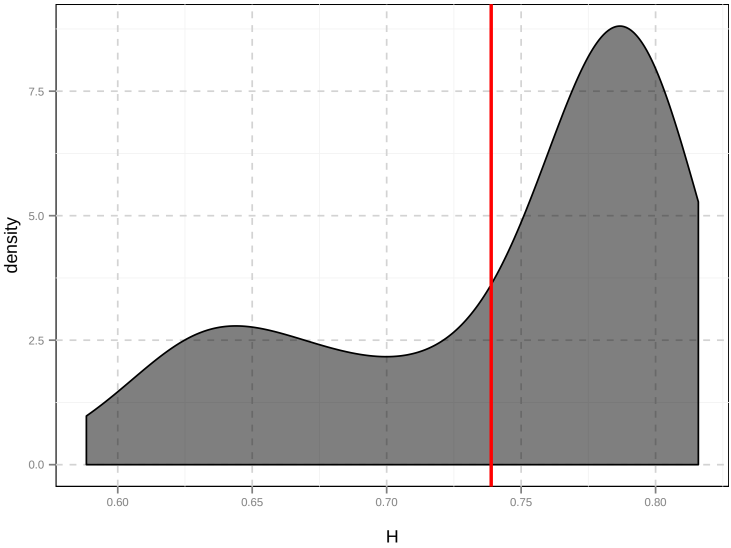

density
30
20
10
0
0.92
0.94
0.96
AUC

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

Best grouping of amino acids - AUC.

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

Worst grouping of amino acids - AUC.

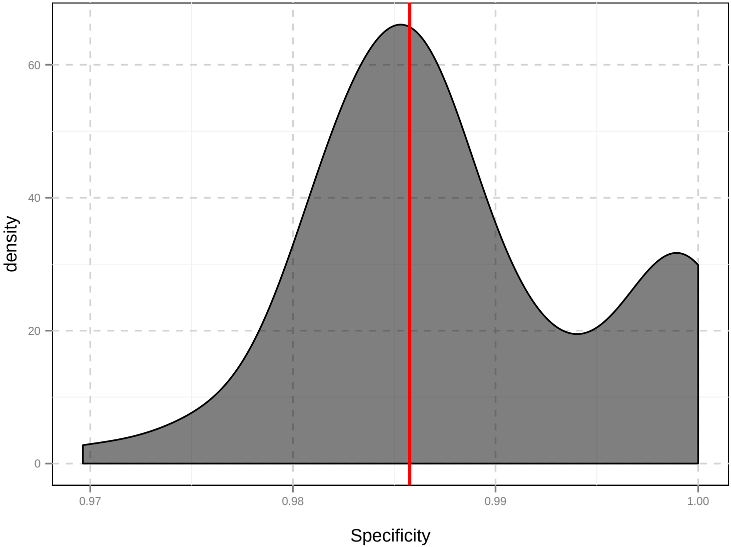

density
7.5
5.0
2.5
0.0
0.60
0.65
0.70
0.75
0.80
H

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

Best grouping of amino acids - H.

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

Worst grouping of amino acids - H.

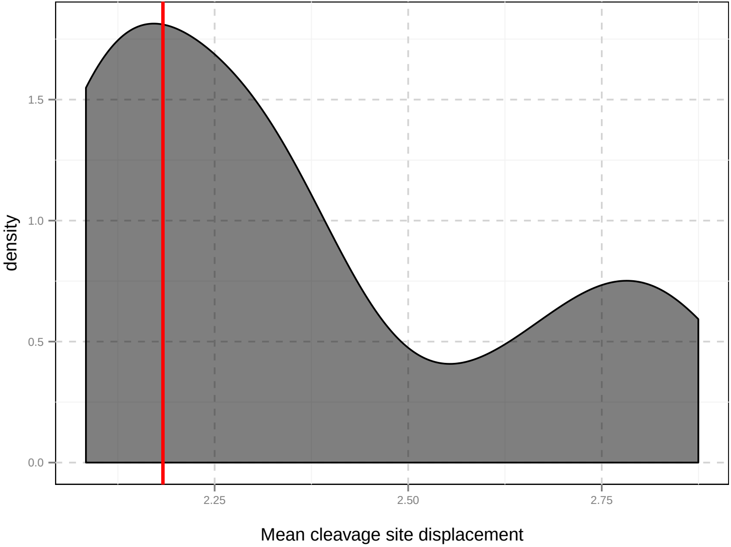

density
60
40
20
0
0.97
0.98
0.99
1.00
Specificity

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

Best grouping of amino acids - specificity.

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

Worst grouping of amino acids - specificity.

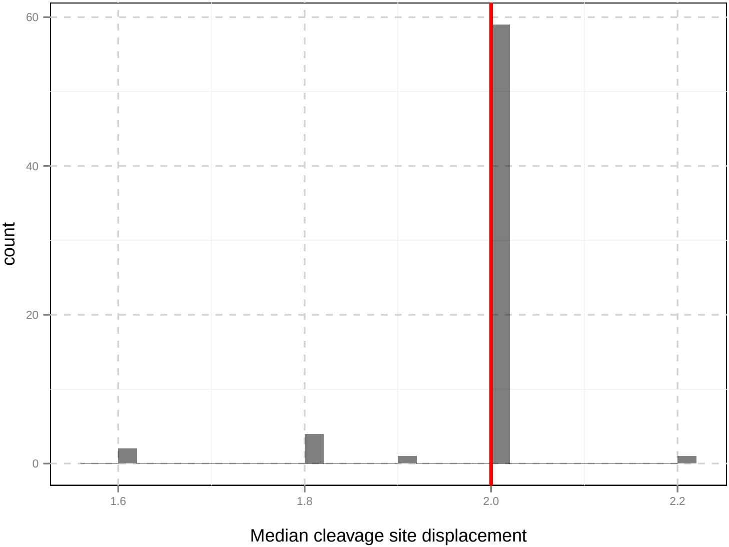

density
1.5
1.0
0.5
0.0
2.25
2.50
2.75
Mean cleavage site displacement

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

Best grouping of amino acids - mean cleavage site displacement.

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

Worst grouping of amino acids - mean cleavage site displacement.


60
40
20
0
count
1.6
1.8
2.0
2.2
Median cleavage site displacement

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

Best (ex aequo) grouping of amino acids - median cleavage site displacement.

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

Best (ex aequo) grouping of amino acids - median cleavage site displacement.

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

Worst grouping of amino acids - median cleavage site displacement.

Outline

1. Motivation
2. Amino acid properties
   - Size
   - Polarity
   - PI
   - Hydrophobicity
   - α-chain
3. Methods
   - Amino acid groupings
   - Cross-validation
4. Cross-validation
   - AUC
   - H
   - Specificity
   - Mean cleavage site displacement
5. Conclusion and perspectives

Grouping on amino acids has strong impact on misclassifications.

Detection of signal peptides:

1. charged amino acids should be grouped separately;
2. aromatic amino acids should not be grouped together;

Detection of cleavage sites:

1. hydrophibic amino acids should be grouped together;