

Amyloid (positive)

PVPMPDLK

Non-amyloid (negative)

PMVMPDKV

Source data: peptides with known amyloidicity status. Overlapping hexamers are marked by horizontal lines.

Extraction of overlapping hexamers with ascribed the amyloid status taken from their source peptide (P-positive, N - negative).

Clusterization of amino acids into an encoding using a combination of various physicochemical properties (PP).

Reduction of the amino acid alphabet in hexamers.

Extraction of n-grams. From each hexamer, we extracted continuous and discontinuous n-grams with the length $n = 1, 2$ or 3 .

Selection of informative n-grams with Quick Permutation Test (QuiPT).

Training of a random forest classifier using the n-grams selected in the previous step.

AA	PP1	PP2	...
A	0.23	0.25	...
C	0.40	0.20	...
D	0.17	0.87	...
E	0.18	0.01	...
G	0.02	1.00	...
...

Ward's
clustering

ID	Amino acids
1	G
2	K, P, R
3	I, L, V
4	F, W, Y
5	A, C, H, M
6	D, E, N, Q, S, T

1	1	1	1
2	1	2	2
3	1	3	3
...

QuiPT

2	2	2	6
3	2	6	2
3	2	2	1
...

