

N-gram analysis of amyloid data

Michał Burdukiewicz¹, Piotr Sobczyk², Paweł Mackiewicz¹ and Małgorzata Kotulska³

¹University of Wrocław, Department of Genomics, Poland

²Wrocław University of Technology, Department of Mathematics, Poland

³Wrocław University of Technology, Department of Biomedical Engineering, Poland

Introduction
Stuff

More stuff
Stuff

AmyLoad database

The sequences used in the study were extracted from AmyLoad database (Wozniak and Kotulska, 2015).

Length	Amyloid (no)	Amyloid (%)	yes	yes (%)
[4,6]	~850	81.03%	~250	59.38%
(6,10]	~120	11.78%	~70	15.44%
(10,15]	~20	2.68%	~40	7.13%
(15,83]	~40	4.5%	~80	18.05%

The height of the bar is equal to the number of sequences within given length interval in the database. The percentage above the bar represent fraction sequences within given length interval among amyloid and non-amyloid sequences.

Bibliography

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2008). AAindex: amino acid index database, progress report 2008. *Nucleic Acids Research*, 36(suppl 1):D202–D205.

Wozniak, P. P. and Kotulska, M. (2014). Characteristics of protein residue-residue contacts and their application in contact prediction. *Journal of Molecular Modeling*, 20(11).

Wozniak, P. P. and Kotulska, M. (2015). AmyLoad: website dedicated to amyloidogenic protein fragments. *Bioinformatics (Oxford, England)*.

Clustering of amino acids

contactivity (Wozniak and Kotulska, 2014)

- 10 physicochemical scale were extracted from AAIndex database (Kawashima et al., 2008) and normalized between normalized between 0 i 1.
- All combinations of characteristics (only one scale per property) were clustered using Euclidean distance and Ward's method.
- Each clustering was divided into 3 to 6 groups.