Prediction of amyloidogenicity based on the n-gram analysis

Michał Burdukiewicz[1], Piotr Sobczyk[2], Stefan Rödiger[3], Anna Duda-Madej[4], Paweł Mackiewicz[1], and Małgorzata Kotulska[5]

[1]**University of Wrocław, Department of Genomics**
[2]**Wrocław University of Science and Technology, Faculty of Pure and Applied Mathematics**
[3]**Brandenburg University of Technology Cottbus-Senftenberg, Institute of Biotechnology**
[4]**Wroclaw Medical University, Department of Microbiology**
[5]**Wrocław University of Science and Technology, Department of Biomedical Engineering, Faculty of Fundamental Problems of Technology**

ABSTRACT

Amyloids are proteins associated with the number of clinical disorders (e.g., Alzheimer's, Creutzfeldt-Jakob's and Huntington's diseases). Despite their diversity, all amyloid proteins can undergo aggregation initiated by 6- to 15-residue segments called hot spots. To find the patterns defining the hot-spots, we trained predictors of amyloidogenicity based on random forests using short subsequences (n-grams) extracted from amyloidogenic and non-amyloidogenic peptides collected in the AmyLoad database. Since the amyloidogenicity may not depend on the exact sequence of amino acids but on more general properties of amino acids, we tested 524,284 reduced amino acid alphabets of different lengths (three to six letters) to find an alphabet providing the best performance of the predictor in cross-validation. The predictor based on this alphabet, called AmyloGram, was benchmarked against the most popular tools for the detection of amyloid peptides using an external data set and obtained the highest values of performance measures (AUC: 0.90, MCC: 0.63). Additionally, our results showed that the amyloidogenicity is strongly correlated with hydrophobicity, a tendency to form β-sheets and rigidity of amino acid residues. After the reduction of amino acids in peptides using the best-performing alphabet, among the most informative n-grams we identified 15 that were already confirmed experimentally. AmyloGram is available as a web-server: www.smorfland.uni.wroc.pl/amylogram/. The code and results are publicly available at: github.com/michbur/prediction_amyloidogenicity_ngram.

Keywords: n-gram, amyloid, prediction, random forest, feature selection

1 INTRODUCTION

Amyloid aggregates have been observed in tissues of people suffering from various diseases, such as: Alzheimer's, Parkinson's, Huntington's and amyotrophic lateral sclerosis, as well as many other conditions. These aggregates were also detected in disorders other than neurological, for example in diabetes of type 2 or certain types of a cataract. Cells in tissues with amyloid oligomers exhibit very high mortality. However, the exact mechanisms of the cytotoxicity have not been discovered. The aggregates are very stable and their dissolution is very difficult. Amyloids are resistant to activity of proteolytic enzymes and chemical compounds due to the specific and highly ordered structure of their steric zipper. However, some strategies to prevent amyloid formation have been proposed, e.g. (Härd and Lendel, 2012).

The aggregation occurs when a cell environment fosters the partial unfolding of protein chains or their fragmentation in such a way that the parts prone to joining with similar protein fragments become exposed. For the majority of proteins, considerable conformational rearrangement must have occurred to initiate the aggregation process. Such changes cannot take place in the typical tightly packed native protein conformation due to the constraints on the tertiary structure. Thus, the formation of the non-native partially unfolded conformation is required to start the aggregation, presumably by enabling specific intermolecular interactions including electrostatic attraction, hydrogen bonding and hydrophobic contacts. This partial unfolding can be influenced by various factors, such as high concentration of proteins, high temperature, low pH, binding metals, or exposition to UV light.

Initially, the resulting molecules form clusters consisting of a few elements, which are called oligomers. Next, they grow into larger aggregates. The aggregation of proteins or their fragments

may lead to amorphous (unstructured) clusters or amyloid (highly ordered) unbranched fibrils. Independently of the protein sequence and its original structure, aggregates always display a common cross-β structure. The distinctive structure of the steric zipper enables the selective detection of amyloids from amorphous aggregates using either a variety of microscopic techniques or fluorescence of probes with which they form compounds.

Currently, it is believed that short peptide sequences of amyloidogenic properties, called hot spots, are responsible for the aggregation of amyloid proteins. Previous studies have suggested that amyloidogenic fragments may have regular characteristics, not only with regard to averaged physicochemical properties of their amino acids, but also the order of amino acids in the sequence. It is important to distinguish between amyloidogenic and amyloid (or amyloidic) peptides, because only the former are capable of initiating the process of aggregation. The latter may consist of amyloidogenic hot-spots as well as other regions that are not directly responsible for the onset of aggregation process, although involved in the final aggregate. Several computational approaches have been proposed to model and predict both kinds of regions. Physics- and chemistry-based models used in FoldAmyloid (Garbuzynskiy et al., 2010) utilize the density of the protein contact sites. Other methods perform threading a peptide on an amyloid fiber backbone, followed by determination of its energy and stability (O'Donnell et al., 2011). Statistical approaches include production of frequency profiles, such as the WALTZ method (Beerten et al., 2015) and machine learning methods, for example those developed in our group (Stanislawski et al., 2013; Gasior and Kotulska, 2014). AGGRESCAN3D was proposed to estimate more accurately aggregation propensity by performing 3D structure based analysis (Zambrano et al., 2015).

The aim of our study is opposite to choose out of thousands of created hot spot models the most appropriate one and from its analysis gain a new insight into the mechanism of amyloidogenicity. To do so, we combined n-gram analysis with the reduction of amino acid alphabet.

In bioinformatics, n-grams (k-mers) are continuous or discontinuous sequences of n elements. Employed as a feature extraction method, n-grams are widely used in various analyses of biological sequences. Our choice of n-grams was driven by their highly interpretable nature. This is a valuable feature because we are interested in identification of motifs that are most relevant to amyloidogenic properties of peptides.

Several studies highlighted that three-dimensional protein structure depends not on the exact sequence of amino acids but also on their general physicochemical properties. Hence, the 20-letter amino acid alphabet reduced into a smaller alphabet (encoding), which represents certain subgroups of aminoacids, can still retain the information about the protein folding (Murphy et al., 2000). Since amyloid aggregates, especially their hot spots regions, have very specific spatial organization, we investigated if these regions can be also described by a shorter amino acid alphabet. Instead of relying on the general similarities of amino acids, we created multiple encodings based on the combinations of various physicochemical properties that might be associated with amyloidogenicity.

To discover amino acid patterns specific for amyloidogenicity, we based our analysis on n-grams, continuous or discontinuous sequences of length n drawn from the encoded peptides. The extraction of n-grams allows the detection of more elaborated motifs, but creates very large features spaces. Henceforth, we used a novel feature selection algorithm, Quick Permutation Test (QuiPT), to select the most informative n-grams.

We used selected n-grams to train a predictor based on the random forest method (Breiman, 2001) to discriminate between amyloidogenic and non-amyloidogenic peptides. We trained the classifier for several iterations on peptides of varying lengths to identify the optimal number of residues which include the information about the occurrence or absence of a hot spot. In the cross-validation setup, we found the encoding associated with the best-performing classifier and its set of informative n-grams. Finally, we benchmarked our best-performing classifier, AmyloGram, on the external data set against state-of-art software tools for prediction of amyloid or amyloidogenic regions.

2 METHODS

2.1 Data set

The data used in the study was extracted from AmyLoad data base (Wozniak and Kotulska, 2015). We obtained 418 amyloidogenic peptides and 1039 non-amyloidogenic peptides.

Sequences shorter than six and longer than 25 amino acid residues (i.e., 8 and 27 sequences, respectively) were removed from the set. The former were too short to be processed in the devised n-gram analysis framework and the latter were too diversified and rare, hampering the proper analysis.

The final data set contained in total 1430 peptides: 397 amyloidogenic and 1033 non-amyloidogenic (and non-amyloid) sequences.

2.2 Encodings of amino acids

The amyloidogenicity of a given peptide may not depend on the exact sequence of amino acids but on their more general properties. To verify this hypothesis, we handpicked 20 different measures from AAIndex data base (Kawashima et al., 2008) describing features important in the amyloidogenicity, such as: size of residues, hydrophobicity, solvent surface area, frequency in β-sheets and contactivity. We preferred the measures that were introduced after 1980 because they were more accurate than the older ones thanks to the technological advancements. The set of selected physicochemical properties was enriched by six measures representing amino acid contact site propensities Wozniak and Kotulska (2014). This gave us 26 features in total .

Since highly correlated measures would create very similar amino acid encodings, we further reduced the number of properties to 17 by selecting measures with the absolute value of Pearson's correlation coefficient smaller than 0.95 (Tab. 1).

Based on that, we created 524,284 encodings with different levels of amino acid alphabet reduction, from three to six groups of amino acid using Ward's clusterization (Joe H. Ward Jr, 1963), which was performed on all combinations of the normalized values of 17 selected physicochemical properties.

The majority of encodings had at least one duplicate. In such a case, only a single representative was included in the cross-validation. After filtering out the duplicates, we obtained 18,535 unique amino acid encodings.

Table 1. Selected 17 physicochemical properties used to create amino acid encodings.

Category	Property
Contactivity	Average flexibility indices (Bhaskaran and Ponnuswamy, 1988)
Contactivity	14 Å contact number (Nishikawa and Ooi, 1986)
Contactivity	Accessible surface area (Radzicka and Wolfenden, 1988)
Contactivity	Buriability (Zhou and Zhou, 2004)
Contactivity	Values of Wc in proteins from class β, cutoff 12 A, separation 5 (Wozniak and Kotulska, 2014)
Contactivity	Values of Wc in proteins from class β, cutoff 12 A, separation 15 (Wozniak and Kotulska, 2014)
β-frequency	Average relative probability of inner beta-sheet (Kanehisa and Tsong, 1980)
β-frequency	Relative frequency in β-sheet (Prabhakaran, 1990)
β-frequency	Thermodynamic β-sheet propensity (Kim and Berg, 1993)
Hydrophobicity	Hydrophobicity index (Argos et al., 1982)
Hydrophobicity	Optimal matching hydrophobicity (Sweet and Eisenberg, 1983)
Hydrophobicity	Hydrophobicity-related index (Kidera et al., 1985)
Hydrophobicity	Scaled side chain hydrophobicity values (Black and Mould, 1991)
Polarity	Polarizability parameter (Charton and Charton, 1982)
Polarity	Mean polarity (Radzicka and Wolfenden, 1988)
Size	Average volumes of residues (Pontius et al., 1996)
Stability	Side-chain contribution to protein stability (kJ/mol) (Takano and Yutani, 2001)

We evaluated advantages of the proposed method for amino acids encoding by adding two standard encodings: (1) ADEGHKNPQRST, C, FY, ILMV, W (Kosiol et al., 2004) and (2) AG, C, DEKNPQRST, FILMVWY, H (Melo and Marti-Renom, 2006), to check if the process of amyloidogenicity does require groupings different from more general amino acid classifications . We also added the full (unreduced) amino acid alphabet to evaluate potential benefits of the alphabet reduction.

2.3 Training sets

In the initial phase, we extracted overlapping hexamers from all peptides. Each hexamer was tagged with one of two etiquettes: amyloid (positive, i.e. originating from an amyloid peptide) or non-amyloid (negative, i.e. originating from a non-amyloid peptide). The etiquette ascribed to the hexamer was based on the amyloid propensity of its source peptide (Fig. 1 A and B). The hexapeptides constituted our training dataset.

Note that amyloid and non-amyloid elements of the set are not necessarily amyloidogenic or non-amyloidogenic, respectively. Hence, assuming that only a short part of the sequence in longer amyloids is responsible for amyloidogenicity, our method might result in many false positives in the training data set and in consequence yield inaccurate predictions as it was evaluated elsewhere (Kotulska and Unold, 2013). To diminish this problem and facilitate the extraction of hot spots, we restricted the maximum length of peptides in the training data set to fifteen amino acids. This

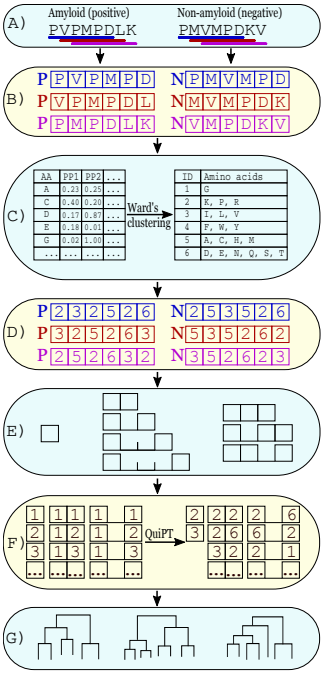

Figure 1. The scheme of n-gram extraction from studied peptide sequences. A) Source data: peptides with known amyloidicity status and indicated overlapping hexamers. B) Extraction of the overlapping hexamers with ascribed the amyloid status taken from their source peptide (P-positive, N - negative). C) Clusterization of amino acids into an encoding using a combination of various physicochemical properties (PP). D) Reduction of the amino acid alphabet in hexamers. E) Extraction of n-grams. From each hexamer, we extracted continuous n-grams with the length n = 1, 2 or 3. In addition, we selected gapped 2-grams with a gap of the length from 1 to 3 residues and gapped 3-grams with a single gap between the first and the second or the second and the third element of the n-gram. F) Selection of informative n-grams with Quick Permutation Test (QuiPT). G) Training of a random forest classifier using the n-grams selected in the previous step.

procedure should eliminate the problem of false negatives and reduce the number of false positives. Moreover, we expect that this influence of false positives would be naturally eliminated or significantly reduced from the pattern finally found in further steps of our method. On the other hand, allowing this ambiguity included many hexamers of potentially high amyloidogenicity, whose propensity has not been experimentally proven.

To further study the problem of the amyloidogenicity signal length, we created three training sets with the sequences of varying lengths (Tab. 2). The smallest data set contained only the sequences of length 6. Assuming that the minimum length of the amyloidogenicity signal is the six residues, we can expect no false positive hexamers in this set. Moreover, we created two training sets with the progressively more liberal limit of the maximum sequence length (6-10 residues and 6-15 residues).

From each hexamer we extracted encoded n-grams (Fig. 1 D and E) with the length of 1, 2 and 3. In the case of 2- and 3-grams, we separately analyzed continuous and gapped n-grams. For 2-grams, we considered n-grams with the gap of the length from 1 to 3, whereas the 3-grams could contain a single gap between the first and the second or the second and the third position (see Fig. 1). The total number of n-grams depends on the the length of the encoding and is equal to 120, 260, 480 and 798 for encodings of length 3, 4, 5 and 6, respectively.

2.4 Quick Permutation Test (QuiPT)

The permutation tests are commonly used for filtering important n-grams. However, they are computationally expensive and, as a result, they often become one of the most limiting factors in these kinds of analyses. Therefore, we developed the Quick Permutation Test which effectively filters n-gram features, without performing a huge number of permutations, using the information gain (mutual information) as the criterion of the importance of a specific n-gram. We used it to select the most discriminating n-grams extracted from the hexamers of the training data set. The counts of n-grams were binarized (1 if n-gram was present, 0 if absent). Only n-grams with the p-value smaller than 0.05 were assumed to be informative (Fig. 1 F).

Table 2. Characteristics of training and test data sets used in the cross-validation.

Set	Sequence length	Status	Sequences	Hexamers
Training	6	Non-amyloid	841	841
		Amyloid	247	247
	[6,10]	Non-amyloid	964	1412
		Amyloid	312	475
	[6,15]	Non-amyloid	992	1653
		Amyloid	342	720
Test	6	Non-amyloid	841	841
		Amyloid	247	247
	[6,10]	Non-amyloid	123	571
		Amyloid	65	228
	[11,15]	Non-amyloid	28	241
		Amyloid	30	245
	[16,25]	Non-amyloid	41	571
		Amyloid	55	778

Let us consider a contingency table for a target y and a feature x. For example, the entry $n_{1,0}$ is the number of cases when the target is 1 and the feature is 0.

target / feature	1	0	total
1	$n_{1,1}$	$n_{1,0}$	$n_{1,\bullet}$
0	$n_{0,1}$	$n_{0,0}$	$n_{0,\bullet}$
total	$n_{\bullet,1}$	$n_{\bullet,0}$	n

Under the hypothesis that x an y are independent, the probability of observing such a contingency table is given by the multinomial distribution. The idea of the permutation test is to reshuffle labels of features and targets, while keeping the fixed total number of positives for features and targets . When we impose this constraint on the multinomial distribution, then the probability of occurrence for a given contingency table depends only on one entry, i.e., $n_{1,1}$, which is fairly easy to compute. After computing Information Gain (IG) for each possible value of $n_{1,1} \in [0, \min(n_{\bullet,1}; n_{1,\bullet})]$, we get the distribution of Information Gain under the hypothesis that the target and feature are independent. We reject the null hypothesis of independence, when IG for the tested feature is above the required quantile from the IG distribution.

The analytic formula for the distribution enables to perform the permutation test much quicker. Furthermore, we get exact quantiles even for extreme tails of the distribution, which is not guaranteed by the random permutations. For example, for the test at the level $\alpha = 10^{-8}$, which can often occur in the corrections for multiple testing, the standard deviation of quantile estimate in the permutation test, $\frac{p(1-p)}{m}$, is roughly equal to α itself even for a huge number of permutations like $m = 10^8$.

In the context of n-gram data, we can further speed up our algorithm. Note that test statistics depends only on $n_{\cdot,1}$, i.e., the number of positive cases in the feature when the target y is common for testing all n-gram features . Although we test millions of features, there are only a few distributions that we need to compute because the usual number of positives in n-gram feature is small. We take advantage of this fact and we compute quantiles only for the handful of distributions. Therefore complexity of our algorithm is roughly equal to $O(n \cdot p)$.

Lastly, let us point out that QuiPT is very similar to Fisher's exact test. From the derivation provided in, e.g. (Lehmann and Romano, 2008), it becomes obvious that QuiPT is heuristics for an unsolved problem of a two-tailed Fisher's exact test. In this heuristics, the extremity of a contingency table is defined by its information gain.

2.5 Random forest classifier

During the learning stage, random forest classifier was trained on the binary n-gram data drawn from the overlapping hexamers, considering only n-grams selected by QuiPT (Fig. 1 G). We grown the forest keeping the default number of tree (500) and the default number of variables to possibly split at in each node (the rounded down square root of the total number of variables). To speed up the computation, we used the fastest implementation of random forest in **R**, ranger package (Wright and Ziegler, 2015).

The predictor separately considered all hexamers coming from a single peptide. If at least one hexamer extracted from a peptide was assessed as amyloidogenic, the whole sequence was denoted

as amyloid. Otherwise, the peptide was classified as non-amyloid. The results were next compared with the known etiquette of the peptides to compute the performance measures.

2.6 Cross-validation and selection of the best-performing encoding

The encoding yielding classifier with the best ability to correctly predict amyloidogenicity of peptides was chosen during the five-fold cross-validation. Since the random assignment of peptides to subsamples in cross-validation may result in the uneven number of hexamers in the subsamples (longer peptides yield more hexamers than shorter ones), we repeated the cross-validation fifteen times for each classifier to obtain more precise estimates of performance measures. We considered three length ranges of sequences in the training sets: 6, 6-10 and 6-15 residues, to evaluate if our classifiers are able to use decision rules extracted from sequences of a different length to correctly classify longer or shorter sequences.

To choose the most adequate amino acid encoding, we ranked the values of the Area Under the Curve (AUC) for each particular classifier (assuming the rank 1 for the best AUC, rank 2 for the second best AUC and so on) and various ranges of the sequence length in the test data set. The encoding with the lowest sum of ranks from all sequence length categories was selected as the best one. For this encoding, we chose the range of the peptides length in the training set that provided the best AUC in the cross-validation.

2.7 Benchmark of AmyloGram

The best-performing encoding that had been chosen during the cross-validation was later used to train AmyloGram, n-gram based predictor of peptide amyloidogenicity.

To compare the performance of AmyloGram and other predictors of amyloids, we used external data set *pep424* (Walsh et al., 2014). Since some peptides were common for both *pep424* and AmyLoad, we removed them from the training data set. After the purification, the training data set for the benchmark consisted of 269 positive sequences and 746 negative sequences, all longer than five and shorter than fifteen residues. Aside from the removal of sequences, the training set of AmyloGram was identical to the training classifiers during the cross-validation. The parameters of QuiPT and random forest algorithms were kept the same.

We removed peptides shorter than five amino acids from the *pep424* data set as our model of amyloidogenicity assumes the minimal length of six residues. Such change should not have affect the outcome of the comparison because only five sequences were eliminated (around 1% of the original data set). Beside the classifier based on the reduced amino acid alphabet, we also benchmarked three predictors based on the full 20-amino acid alphabet learned on n-grams extracted from sequences of different length ranges to separately assess the benefit of using only the n-gram analysis without the reduction of amino acid alphabet.

3 RESULTS AND DISCUSSION

3.1 Performance of the best encoding

The predictor based on the chosen encoding had the AUC always in the fourth quartile of all AUC values. (Fig. 2). It reached the highest AUC (0.8667) in classification of the shortest sequences (with the length of 6 residues) when the training set also consisted of the sequences of the same length. It results most probably from the pattern homogeneity of the short peptides. Nevertheless, the results were quite comparable even if the training set contained longer sequences. On average, the majority of encodings behaved like the selected encoding and also received higher AUC when tested on the shorter sequences than for the longer ones. For the best-performing encoding, the most problematic was the correct prediction of the amyloidogenicity in the longest peptides, ranging from 16 to 25 residues, when the algorithm was trained also on longer peptides (6-10 and 6-15 data sets). Here the AUC value did not exceed 0.77.

The weakest performance measured by AUC for sequences longer than 15 amino acids results from difficulties in extraction of the important discriminating features related with more complex organization of longer amylogenic peptides. In such peptides, only a very specific region of residues might be responsible for the creation of harmful aggregates. In this case, when overlapping hexamers are extracted, only part of them may carry the true signal of amyloidogenicity but all of them are marked as amyloids. However, despite this problem, the overall prediction of classifiers learned on the long sequences was still adequate, with the mean values of AUC higher than 0.7 for every testing set.

We also evaluated classifiers based on the full (i.e., unreduced) amino acid alphabet. In most cases, they were also placed in the fourth quartile of the AUC values (Fig. 2). Nevertheless, they

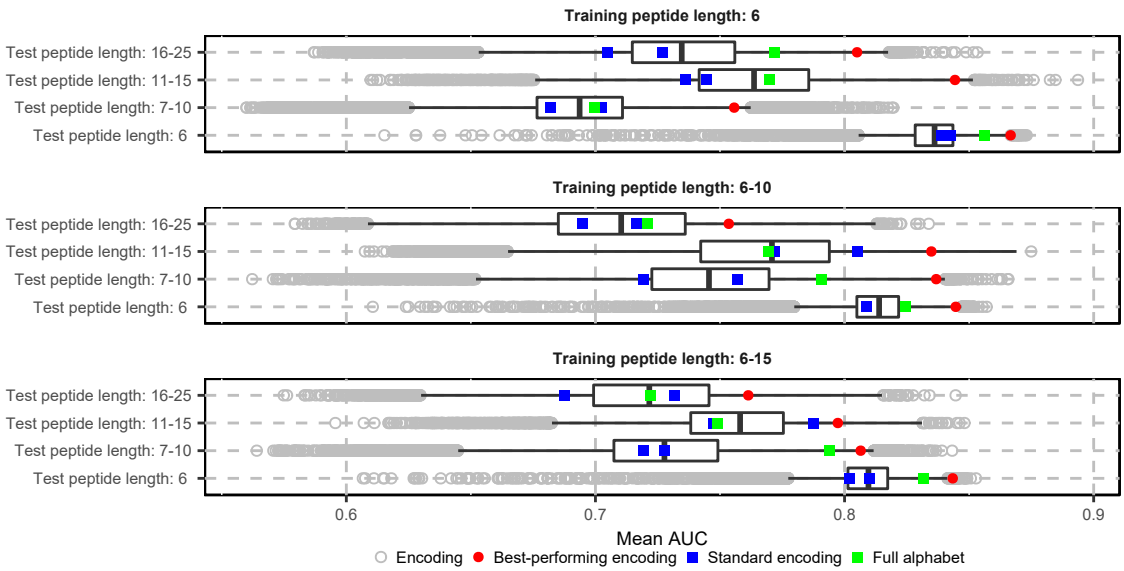

Figure 2. Distribution of mean AUC values of classifiers with various encodings for every possible combination of training and testing data set including different lengths of sequences. The left and right hinges of boxes correspond to the 0.25 and 0.75 quartiles. The bar inside the box represents the median. The gray circles correspond to the encodings with the AUC outside the 0.95 confidence interval.

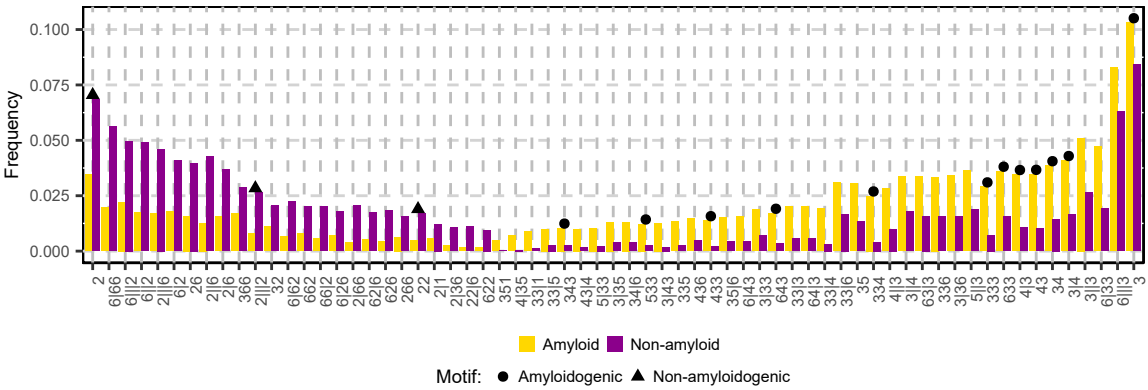

Figure 3. The frequency of important n-grams used by the best-performing classifier in amyloid and non-amyloid sequences. The elements of n-grams are amino acids encoded using the best-performing reduced amino acid alphabet. A vertical bar represents a gap in a n-gram between its elements. The frequency was computed using the total number of occurrences divided by the number of possible n-grams of their length. Dots and triangles denote n-grams occurring in motifs found in respectively amyloidogenic and non-amyloidogenic sequences (López de la Paz and Serrano, 2004).

never predicted amyloidogenicity better than the best classifier based on the reduced alphabet. It indicates that despite an abundance of information present in the full alphabet, the selected encoding allows for better generalization of the prediction rules and in the consequence the better performance.

3.2 The best-performing encoding and important n-grams

In total, eleven combinations of physicochemical properties created the best performing encoding. Only four features appeared in all combinations: hydrophobicity index (Argos et al., 1982), average flexibility indices (Bhaskaran and Ponnuswamy, 1988), polarizability parameter (Charton and Charton, 1982) and thermodynamic β-sheet propensity (Kim and Berg, 1993).

The best encoding chosen in the analysis consists of six amino acid subgroups, which are characterized by distinct and specific properties: 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. The 3rd subgroup contains strongly hydrophobic amino acids. In the 4th subgroup, the amino acids show also aromatic properties. On the other hand, the most hydrophilic amino acids are in the 2nd and 6th subgroups . The former includes two strongly basic amino acids,

whereas the latter two acidic and four polar residues. The first subgroup includes only glycine, which is the smallest amino acid and the most flexible. By average, quite flexible amino acids are also present in the second subgroup, whereas the least flexible amino acids are in the subgroup 4 and 5. The glycine has also the lowest propensity to form β-sheet and the subgroups 3 and 4 largest.

We found 65 n-grams that had obtained p-values smaller than 0.05 in QuiPT test in all repetitions of cross-validation, regardless of the lengths of sequences in the training set (see Fig. 3). The frequency of the n-grams was computed for all sequences derived from AmyLoad . The n-grams typical of amyloidogenic sequences (with the highest frequency occurrence in amyloids) mostly include highly hydrophobic amino acids with tendency to form β-structures, from subgroups 3 and 4. The n-grams occurring frequently in amyloids have often repeats of amino acids from the subgroup 3, suggesting that the presence of these amino acids in the vicinity might be one of the most effective predictors of amyloidogenicity.

On the other hand, n-grams typical of non-amyloidogenic peptides have mostly amino acids belonging to subgroups 2 and 6. These subgroups include strongly hydrophilic and highly flexible amino acids (K, P, R, D, E), which hamper the formation of β-structures.

Out of 65 the most informative n-grams, 15 (23%) were also found in the motifs validated experimentally for amyloidogenic and non-amyloidogenic peptides (López de la Paz and Serrano, 2004). The peptides used in this study are included in the AmyLoad data base, thus n-gram analysis is at least partially able to find the patterns in validated sequences.

3.3 Benchmark of AmyloGram

Table 3. Results of benchmark on *pep424* data set for PASTA2, FoldAmyloid, AmyloGram and its version learned on n-grams extracted for full amino acid alphabet from the sequences of the lengths specified in the brackets.

Classifier	AUC	MCC	Sensitivity	Specificity
AmyloGram	**0.8972**	**0.6307**	0.8658	0.7889
full alphabet (6)	0.8411	0.5427	0.4966	**0.9593**
full alphabet (6-10)	0.8581	0.5698	0.7517	0.8259
full alphabet (6-15)	0.8610	0.5490	0.8188	0.7519
PASTA2	0.8550	0.4291	0.3826	0.9519
FoldAmyloid	0.7351	0.4526	0.7517	0.7185
APPNN	0.8343	0.5823	**0.8859**	0.7222

The benchmark covered AmyloGram as well as the three peer-reviewed predictors of amyloidogenicity: PASTA2 (Walsh et al., 2014), FoldAmyloid (Garbuzynskiy et al., 2010) and APPNN (Família et al., 2015). Other classifiers published before were not included in the benchmark because their performance on *pep424* data set is already known and lower than the performance of PASTA2 and FoldAmyloid (Walsh et al., 2014).

We analyzed AUC, Matthew's Correlation Coefficient (MCC), sensitivity and specificity (see Tab. 3). We used default settings for FoldAmyloid, PASTA2 evaluated the input data in the 'Peptides' mode.

Since PASTA2 does not return a probability of belonging to a specific category, we normalized the output data to compute the AUC values. The advised energy threshold (-5) was also normalized in the same manner and used as cut-off in computations of specificity, sensitivity and MCC. The resulting value of specificity 0.9519 is close to the value provided by its authors (0.95) and assures correctness of our computations. For other classifiers, we assumed a 0.5 cut-off.

In the case of the studied data set, the n-gram extraction method appeared efficient enough to produce classifiers able to outperformed the other published methods. AmyloGram showed the highest AUC and MCC among all tested classifiers. It had lower specificity than two n-gram classifiers trained on the full alphabet but still outperformed two other published method in this category. It is important to highlight that AmyloGram is the most balanced tool among all analyzed classifiers, having the best specificity/sensitivity trade-off, as indicated by the value of MCC.

The specificity of AmyloGram is lower than the specificity of PASTA2 but it is a consequence of the usage of the threshold value optimized for 0.95 specificity for the latter. If we assume for the AmyloGram the same threshold for the specificity, our classifier will still have a higher sensitivity (0.5518) than PASTA2. Therefore, if we assume such thresholds to both predictors, they will detect true non-amyloids with the same specificity but AmyloGram will predict more true amyloids.

Two of the three classifiers trained on n-grams using the full alphabet had also AUCs higher than PASTA2 and all three were more successful than FoldAmyloid as well as APPNN. They also maintained the high specificity as seen previously during cross-validation. However, they generally performed worse than the classifier based on the reduced amino acid alphabet.

Among all considered predictors of amyloidogenicity, APPNN had the highest sensitivity. Nevertheless, its AUC was worse than AUCs of all n-gram-based predictors, as well as PASTA2 - indicating lower overall performance.

AmyloGram is trained to predict amyloidogenic, not amyloidic regions. Hence, we did not tested it on *reg33* data set, which is commonly used to evaluate the amyloid propensity of the full peptide (Tsolis et al., 2013).

4 CONCLUSION

Thanks to the reduction of the amino acid alphabet and description of peptides by short sub-sequences (n-grams) , we were able to create the efficient predictor of amyloidogenic sequences called AmyloGram. One of the strengths of the assumed approach is its highly interpretable outcome, because our methods provide explicitly short motifs relevant to amyloidogenicity of peptides and discriminating amyloids from non-amyloids. 65 important n-grams revealed that mostly alifatic and nonpolar amino acids (isoleucine, leucine and valine), together with aromatic and also hydrophobic amino acids (phenylalanine, tyrosine, tryptophan), are good predictors of amyloid peptides.

Our studies confirm that the most important physicochemical properties associated with amyloidogenicity are hydrophobicity and tendency to forming β-sheets. We additionally discovered that amino acid flexibility can also sufficiently discriminate amyloid and non-amyloid peptides. The aggregating peptides tend to be characterized by more rigid chains.

Our findings can be helpful in understanding the process of amyloid aggregation and recognition of peptides susceptible to the formation of amyloid aggregates involved in various diseases. Moreover, they might be employed in the creation of synthetic amyloid peptides. We anticipate that the described workflow is versatile enough to be applied in other areas of protein function prediction.

FUNDING AND AVAILABILITY

Computations were carried out in Wroclaw Center for Networking and Supercomputing (http://www.wcss.pl) and funded by the institutional grant No. 347. This research was funded by the KNOW Consortium and National Science Center (2015/17/N/NZ2/01845).

Our software is accessible as the web server under address www.smorfland.uni.wroc.pl/amylogram/.

The code and results are publicly available at: github.com/michbur/prediction_amyloidogenicity_ngram.

REFERENCES

Argos, P., Rao, J. K., and Hargrave, P. A. (1982). Structural prediction of membrane-bound proteins. *European journal of biochemistry / FEBS*, 128(2-3):565–575.

Beerten, J., Van Durme, J., Gallardo, R., Capriotti, E., Serpell, L., Rousseau, F., and Schymkowitz, J. (2015). WALTZ-DB: a benchmark database of amyloidogenic hexapeptides. *Bioinformatics (Oxford, England)*, 31(10):1698–1700.

Bhaskaran, R. and Ponnuswamy, P. (1988). Positional flexibilities of amino acid residues in globular proteins. *International Journal of Peptide and Protein Research*, 32(4):241–255.

Black, S. D. and Mould, D. R. (1991). Development of hydrophobicity parameters to analyze proteins which bear post- or cotranslational modifications. *Analytical Biochemistry*, 193(1):72–82.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Charton, M. and Charton, B. I. (1982). The structural dependence of amino acid hydrophobicity parameters. *Journal of Theoretical Biology*, 99(4):629–644.

Família, C., Dennison, S. R., Quintas, A., and Phoenix, D. A. (2015). Prediction of Peptide and Protein Propensity for Amyloid Formation. *PLOS ONE*, 10(8):e0134679.

Garbuzynskiy, S. O., Lobanov, M. Y., and Galzitskaya, O. V. (2010). FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics (Oxford, England)*, 26(3):326–332.

Gasior, P. and Kotulska, M. (2014). FISH Amyloid – a new method for finding amyloidogenic segments in proteins based on site specific co-occurence of aminoacids. *BMC Bioinformatics*, 15(1):54.

Härd, T. and Lendel, C. (2012). Inhibition of Amyloid Formation. *Journal of Molecular Biology*, 421(4–5):441–465.

Joe H. Ward Jr (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244.

Kanehisa, M. I. and Tsong, T. Y. (1980). Local hydrophobicity stabilizes secondary structures in proteins. *Biopolymers*, 19(9):1617–1628.

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.

Kidera, A., Konishi, Y., Oka, M., Ooi, T., and Scheraga, H. A. (1985). Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *Journal of Protein Chemistry*, 4(1):23–55.

Kim, C. A. and Berg, J. M. (1993). Thermodynamic beta-sheet propensities measured using a zinc-finger host peptide. *Nature*, 362(6417):267–270.

Kosiol, C., Goldman, N., and Buttimore, N. H. (2004). A new criterion and method for amino acid classification. *Journal of Theoretical Biology*, 228(1):97–106.

Kotulska, M. and Unold, O. (2013). On the amyloid datasets used for training PAFIG how (not) to extend the experimental dataset of hexapeptides. *BMC Bioinformatics*, 14:351.

Lehmann, E. L. and Romano, J. P. (2008). *Testing Statistical Hypotheses*. Springer New York.

López de la Paz, M. and Serrano, L. (2004). Sequence determinants of amyloid fibril formation. *Proceedings of the National Academy of Sciences of the United States of America*, 101(1):87–92.

Melo, F. and Marti-Renom, M. A. (2006). Accuracy of sequence alignment and fold assessment using reduced amino acid alphabets. *Proteins*, 63(4):986–995.

Murphy, L. R., Wallqvist, A., and Levy, R. M. (2000). Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Engineering*, 13(3):149–152.

Nishikawa, K. and Ooi, T. (1986). Radial locations of amino acid residues in a globular protein: correlation with the sequence. *Journal of Biochemistry*, 100(4):1043–1047.

O'Donnell, C. W., Waldispühl, J., Lis, M., Halfmann, R., Devadas, S., Lindquist, S., and Berger, B. (2011). A method for probing the mutational landscape of amyloid structure. *Bioinformatics*, 27(13):i34–i42.

Pontius, J., Richelle, J., and Wodak, S. J. (1996). Deviations from Standard Atomic Volumes as a Quality Measure for Protein Crystal Structures. *Journal of Molecular Biology*, 264(1):121–136.

Prabhakaran, M. (1990). The distribution of physical, chemical and conformational properties in signal and nascent peptides. *The Biochemical Journal*, 269(3):691–696.

Radzicka, A. and Wolfenden, R. (1988). Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry*, 27(5):1664–1670.

Stanislawski, J., Kotulska, M., and Unold, O. (2013). Machine learning methods can replace 3d profile method in classification of amyloidogenic hexapeptides. *BMC Bioinformatics*, 14:21.

Sweet, R. M. and Eisenberg, D. (1983). Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *Journal of Molecular Biology*, 171(4):479–488.

Takano, K. and Yutani, K. (2001). A new scale for side-chain contribution to protein stability based on the empirical stability analysis of mutant proteins. *Protein Engineering*, 14(8):525–528.

Tsolis, A. C., Papandreou, N. C., Iconomidou, V. A., and Hamodrakas, S. J. (2013). A Consensus Method for the Prediction of 'Aggregation-Prone' Peptides in Globular Proteins. *PLOS ONE*, 8(1):e54175.

Walsh, I., Seno, F., Tosatto, S. C. E., and Trovato, A. (2014). PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Research*, page gku399.

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)*.

Wright, M. N. and Ziegler, A. (2015). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *arXiv:1508.04409 [stat]*. arXiv: 1508.04409.

Zambrano, R., Jamroz, M., Szczasiuk, A., Pujols, J., Kmiecik, S., and Ventura, S. (2015). AGGRESCAN3d (A3d): server for prediction of aggregation properties of protein structures. *Nucleic Acids Research*, page gkv359.

Zhou, H. and Zhou, Y. (2004). Quantifying the effect of burial of amino acid residues on protein stability. *Proteins*, 54(2):315–322.