listofformat subrefformat
[subfloat]font=footnotesize,
labelformat=parens,labelsep=space,
listofformat=subparens,subrefformat=subsimple
[subfloat] [subfigure] [subtable]

Published online 31 July 2009 *Nucleic Acids Research, 2009, Vol. 37, No. 12* **1–9**
doi:10.1093/nar/gkn000

Prediction of amyloidogenicity based on the n-gram analysis

Michał Burdukiewicz [1], Piotr Sobczyk [2], Paweł Mackiewicz [1] and Małgorzata Kotulska [3] *

[1]University of Wrocław, Department of Genomics, [2]Wrocław University of Science and Technology, Department of Mathematics and [3]Wrocław University of Science and Technology, Department of Biomedical Engineering, Faculty of Fundamental Problems of Technology

Received on XXXXX; revised on XXXXX; accepted on XXXXX

ABSTRACT

Text. Text.

INTRODUCTION

Amyloid aggregates have been observed in tissues of people suffering from Alzheimer's disease, Parkinson's disease, amyotrophic lateral sclerosis and Huntington's disease, as well as many other conditions. They also include diseases other than neurological, for example diabetes type 2 or certain types of a cataract. Cells in tissues with amyloid fibrils exhibit very high mortality. However, the mechanisms of the cytotoxicity have not been discovered. Unfortunately dissolution of the aggregates 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.

The aggregation occurs when a cell environment fosters the partial unfolding of protein chains or their fragmentation, in a way that the parts prone to joining with other protein fragments are 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 of the tertiary structure. Thus, formation of a non-native partially unfolded conformation is required, presumably enabling specific intermolecular interactions, including electrostatic attraction, hydrogen bonding and hydrophobic contacts. This partial unfolding can be influenced by various factors, such as protein high concentration, 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. 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) can be responsible for 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. There have been attempts to predict the sequence of such peptides by computational modelling. Physics and chemistry based models have been used, including FoldAmyloid (Garbuzynskiy et al., 2010). This method is based on 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 (Bryan et al., 2012, Goldschmidt et al., 2010, 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, which have been used by our team (Gasior and Kotulska, 2014, Stanislawski et al., 2013). AGGRESCAN3D has been proposed to estimate more accurately aggregation propensity by performing 3D structure based analysis (Zambrano et al., 2015).

In this study we present an n-gram model of amyloidogenic sequences. 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 analysis of biological sequences. Our choice of n-grams was driven by their highly interpretable nature. This is a valuable feature since we are interested in identification of motifs that are most relevant to amyloidogenic properties of peptides. Out of several possible n-grams the most relevant features using the novel feature selection algorithm called Quick Permutation Test (QuiPT).

*To whom correspondence should be addressed. Email: malgorzata.kotulska@pwr.edu.pl

Several studies highlighted that three-dimensional protein structure depends not on the exact sequence of amino acids, but on their general physicochemical properties. Hence, a reduced amino acid alphabet, where a single element represents several amino acids, still retains the information about protein folding (Murphy et al., 2000). Since amyloid aggregates, especially their hot spots regions, have very specific spatial organization, we investigated if it also can be described using a shorter alphabet. Instead of relying on general similarities of amino acids, we created our own reduced amino acid alphabets based on the combinations of various physico-chemical properties that might be associated with amyloidogenicity. Due to that we discovered which of these properties best discriminate between amyloids and non-amyloids.

Our model of amyloidogenicity consist of n-gram datasets extracted from sequences encoded using different reduced amino acid alphabets. The information in models was further used to train predictors of amyloids based on random forest (Breiman, 2001). We trained several iterations of each classifier using peptides of different length to identify the optimal number of residues consisting the information of the hot spot presence or absence. Through the cross-validation of predictors we determined the best-performing classifier, its reduced amino acid alphabet and set of important n-grams.

METHODS

Data set

The data used in the study was extracted from AmyLoad data base (Wozniak and Kotulska, 2015). Aside from eight sequences shorter than five residues that were removed from the final data set, we obtained 418 amyloidogenic sequences and 1039 non-amyloidogenic sequences (1457 peptides in total).

Sequences shorter than 6 amino acids and longer than 25 amino acids were removed from the data set. The former were too short to be processed in the devised analysis framework and the latter were too diversified and rare, hampering the proper analysis.

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

Encodings of amino acids

The amyloidogenicity of a given peptide may not depend on the exact sequence of amino acids, but on its more general properties. To verify this hypothesis, we created 524 284 reduced amino acid alphabets (encodings) with different lengths (from three to six letters) using Ward's clusterization (Joe H. Ward Jr, 1963) on the selected physicochemical properties from AAIndex database (Kawashima et al., 2008). We handpicked several measures belonging to more general categories important in the amyloidogenicity, such as size, hydrophobicity, solvent surface area, frequency in β-sheets and contactivity. As a rule of thumb, we limited it to properties introduced after 1980 when, thanks to the technological advancements, the measurements were more accurate.

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 duplicates, we obtained 18 535 unique encodings.

Category	Property
Contactivity	Average flexibility indices (Bhaskaran and Ponnuswamy, 1988)
Contactivity	14 A 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)

table

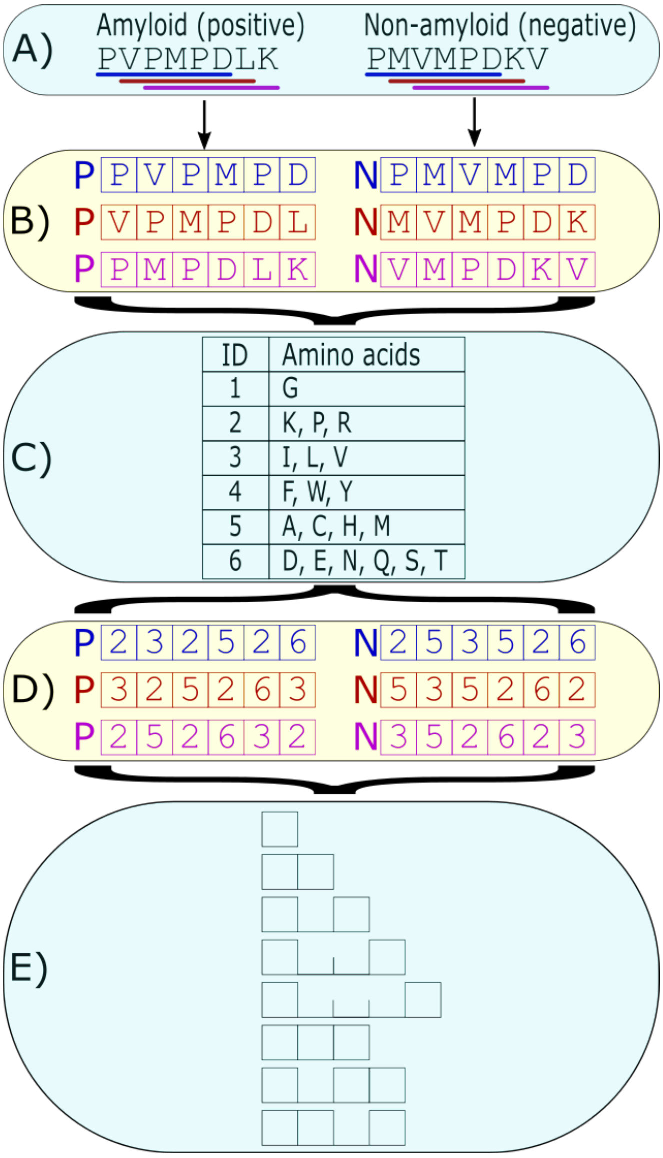
Table 1. Physicochemical properties used in the creation of reduced amino acid alphabets.

Since correlated or, contrarily, uncorrelated measures create very similar encodings, we further reduced the number of properties to 17 by selecting measures with the absolute value of Pearson's correlation coefficient larger than 0.95 for normalized values (Tab. 1).

Training of learners

During the training phase, we extracted overlapping hexamers from each sequence. Each hexamer was tagged with the same etiquette (amyloid/nonamyloid) as the source peptide. For example, the amyloidogenic sequence of length 6 residues yields only one hexamer tagged as "amyloid" and a non-amyloidogenic sequence of 8 residues yields 3 hexamers, all marked as "non-amyloids". (Fig. 1 A and B).

Assuming that in longer amyloids only a short part of the sequence 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



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

figure

Figure 1. AAThe scheme of n-gram extraction. A) Input data - peptides with a known amyloidogenicity status. B) Each peptide sequence was divided into overlapping hexamers. The amyloidogenicity status of a source sequence was used as the amyloidogenicity status of a derived hexamer. C) From each hexamer we extracted continuous 1-, 2- and 3-grams. We selected also gapped 2-grams with the length of gap equal 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.

evaluated elsewhere (Kotulska and Unold, 2013). To diminish this problem and ease the extraction of hot spots, we restricted the maximum length of peptides in training data set to fifteen amino acids.

To study further the problem of the length of an amyloidogenicity signal, we created three learning sets with the sequences of different lengths (Tab. 2). The smallest data set contains only the sequences of length 6. Assuming that the minimum length of the amyloidogenicity signal is six residues, we expect no false positive hexamers. Moreover, we created two training sets with the progressively more liberal

table

Table 2. Sizes of training data sets used in the analysis.

Length range	Status	n (sequences)	n (hexamers)
6	Non-amyloid	841	841
	Amyloid	247	247
6-10	Non-amyloid	123	1412
	Amyloid	65	475
6-15	Non-amyloid	28	1653
	Amyloid	30	720.00

limit of the maximum sequence length (6-10 residues and 6-15 residues).

the extraction of probable hot-spots.

From each hexamer we extracted n-grams of the following lengths: 1, 2 and 3. In the case of 2- and 3-grams, we separately counted both gapped and continuous n-grams. For 2-grams we considered n-grams with gaps of lengths from 1 to 3 and for 3-grams with a single gap between the first and the second or the second and the third element (see Fig. 1).

All n-grams extracted from the hexamers in the training data set were filtered using described below Quick Permutation Test with the information gain (mutual information) as the criterion of the importance of a specific n-gram. In the next step, we used n-grams with the p-value smaller than 0.05 to build a random forest classifier using ranger **R** package (Wright and Ziegler, 2015).

Furthermore, we repeated the procedure described above on two typical reduced alphabets of amino acids derived from the literature to check if the process of amyloidogenicity does require nonstandard groupings of amino acids. We also added the full amino acid alphabet to assess the advantages of the amino acid encoding.

Quick Permutation Test (QuiPT)

The permutation tests, commonly used for filtering important n-grams, are computationally expensive, and, as a result, they often become one of the most limiting factors for these kinds of analysis. The Quick Permutation Test effectively filters n-gram features without performing a huge number of permutations. Let us consider the contingency table for target y and feature x. For example entry n_{10} is the number of cases when target is 1 and 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 permutation test is to reshuffle feature and target labels, while keeping the total number of positives in both of them fixed. When we impose this constraint on the multinomial distribution, then the probability of occurrence for a given contingency table depends only on one entry, say $n_{1,1}$, and 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 hypothesis that target and feature are independent. We reject null hypothesis, when IG for a feature we test is above a required quantile from IG distribution.

Having the analytical formula for the distribution, enables us to perform permutation test much quicker. Furthermore, we get exact quantiles even for extreme tails of the distribution, which is not guaranteed by random permutations. In fact, imagine performing the test with $\alpha = 10^{-8}$, which is not an uncommon value, e.g. when one adjusts for multiple testing. Even for a huge number of permutations like $m = 10^8$, the standard deviation of quantile estimate in permutation test, $\frac{p(1-p)}{m}$, is roughly equal to α itself.

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

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 a heuristics for an unsolved problem of a two-tailed Fisher's exact test. In this heuristics, extremity of a contingency table, is defined by its Information Gain.

Cross-validation and selection of the best-performing encoding

The ability to correctly predict amyloidogenicity was assessed during the five-fold cross-validation. The peptides were assigned randomly to subsamples. Since this approach may result in the uneven number of hexamers between subsamples (single peptide longer than six resides yields more than single hexamer), repeated the cross-validation fifteen times for each classifier to obtain more precise estimates of performance measures for each classifier.

During the testing phase, if at least one hexamer extracted from a peptide was assesed as amyloidogenic, the whole sequence was denoted as amyloid. Otherwise, the peptide was classified as non-amyloid. The results were later confronted with the known etiquettes of the peptides to compute the performance measures.

To evaluate if our classifiers are able to use decision rules extracted from sequences of given length to correctly classify longer or shorter sequences, we calculate performance measures separately for four ranges of lengths of sequences: 6, 7-10, 11-15 and 16-25. The number of sequences from the given length range was roughly comparable between folds of cross-validation.

To choose the most adequate reduced amino acid alphabet, we ranked the values of Area under Curve - AUC (with rank 1 for the best AUC, rank 2 for the second best AUC and so on) for each range of sequence length in the testing 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 choose the range of peptides length in the training set providing the best AUC in cross-validation.

Encoding distance

The encoding distance is a measure defining the similarity between two encodings. It has zero value for identical encodings and grows with the differences between encodings. It was introduced to verify if the reduced alphabets very similar to the best-performing encoding will also have good prediction performance.

We define the encoding distance as the minimum number of amino acids that have to be moved between subgroups of encoding to make a identical to b (the order of subgroups in the encoding and amino acids in a group is unimportant). This measure is further scaled by a factor reflecting how much moving amino acids between groups altered mean group properties.

To compute the scale factor s for the encoding distance between encoding a with n subgroups and encoding b with m subgroups we first calculate p_i and p_j, mean values of physicochemical properties of all amino acids separately for each subgroup. The factor between a and b is equal to:

$$s_{ab} = \sum_{i=1}^{n} \left(\min_{j=1,\ldots,m} \left(\sqrt{\sum^{l} p_i^2} - \sqrt{\sum^{l} p_j^2} \right) \right)$$

where l is equal to the number of physicochemical properties of concern.

Benchmark of AmyloGram

The best-performing reduced amino acid alphabet chosen during the cross-validation was later used to train AmyloGram, n-gram based predictor of amyloidogenicity.

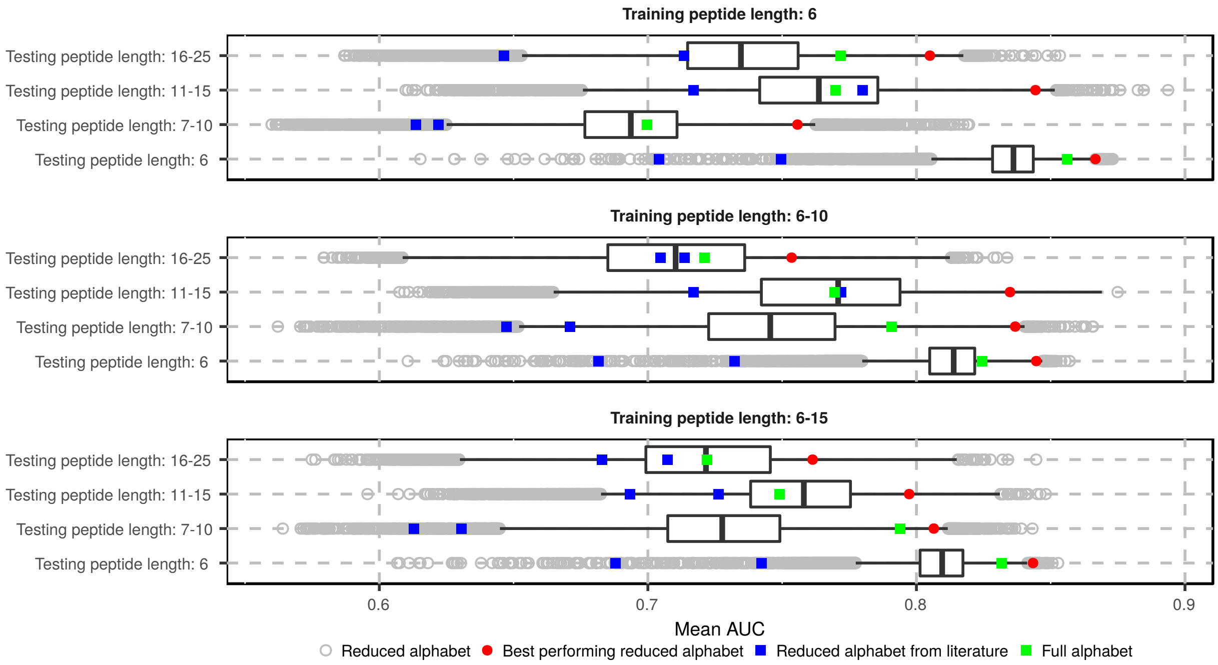
We used *pep424* data set (Walsh et al., 2014) to compare the performance of AmyloGram and other predictors of amyloidogenicity. Since the model of AmyloGram does not cover peptides shorter than five amino acids, we removed them from the total benchmark data set. It should have not affect the comparison as only five sequences were eliminated (around 1% of the original data set). Additionally, we also benchmarked three predictors learned on the n-grams extracted from sequences of different length ranges without any amino acid encoding.

All benchmarked classifiers were trained on sequences used during the cross-validation. Since some peptides were common for both *pep424* and AmyLoad, we removed them from the training data set. After purification, the learning data set had 269 positive sequences and 746 negative sequences longer than five residues and shorter than fifteen residues. Aside from the preparation of the training data, we exactly repeated the procedure of n-gram extraction as described above (see Fig. 1).

RESULTS AND DISCUSSION

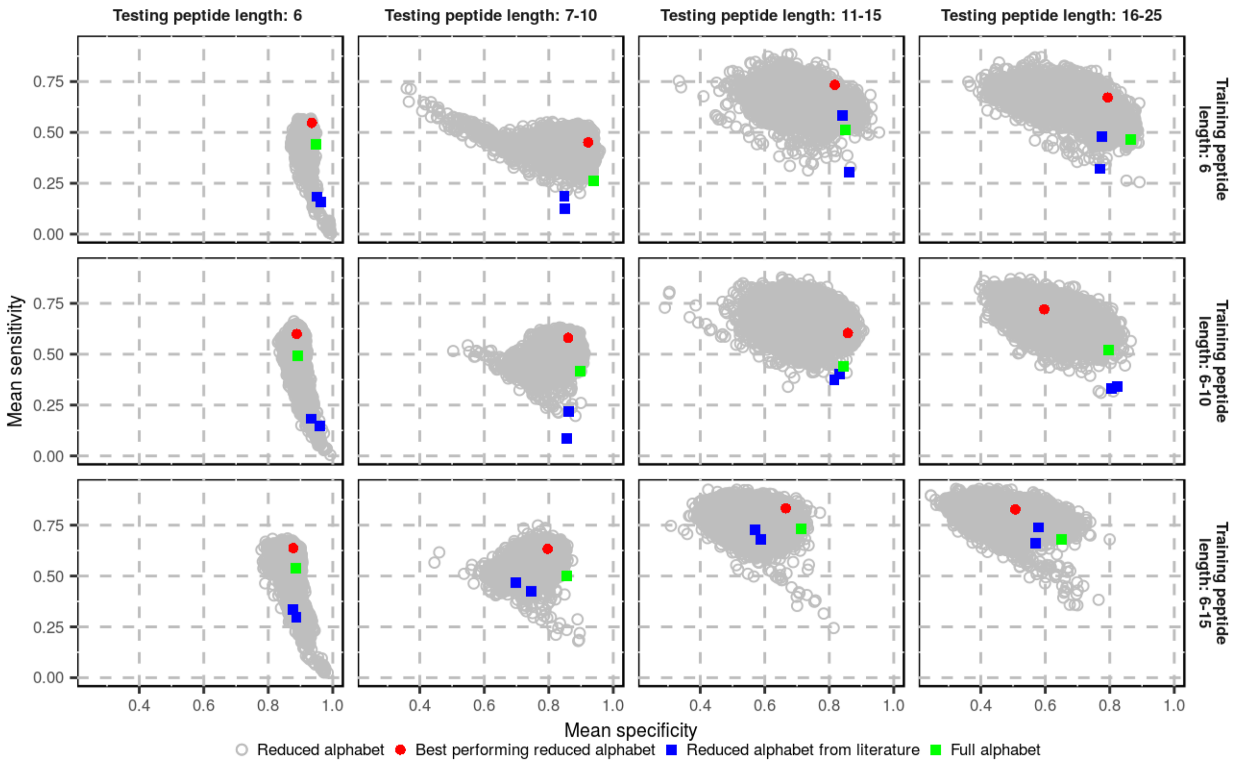
Selection of the best-performing encoding

The selected encoding performed better than other reduced alphabets considering all sequence length ranges in training and testing data set. It had the AUC always in the fourth quartile (Fig. 2). For the best-performing encoding the most problematic was correct prediction of the amyloidogenicity in



figure

Figure 2. Distribution of AUC values of all classifiers based on reduced amino acid alphabets for every possible combination of training and testing data set. for different lengths of sequences in the training and testing data sets. The left and right hinges of boxes correspond to the 0.25 and .75 quartiles. The bar inside the box represents the median. All gray points corresponds to encodings with the AUC outside the 0.95 confidence interval.



figure

Figure 3. Sensitivity and specificity of classifiers in cross-validation for different lengths of sequences in the training and testing data sets. The classifier based on the best-performing encoding always have good specificity and sensitivity.

the longest peptide data set 16-25, not when learned on very short sequences, but for longer ones (6-10 and 6-15). Such a behavior was typical for most of the analyzed encodings.

The chosen encoding reaches the highest values of AUC in the relatively simplest cases of predicting amyloidogenicity, i.e. for the shortest sequences (6 residues). Of course, the decision rules were the easiest to extract when also the learning data set had only sequences of the same lengths, but the results were quite comparable even if the training set contained longer sequences. Also in this situation, the majority of reduced amino acid alphabets behaved like the selected encoding and also had higher AUC than in other scenarios.

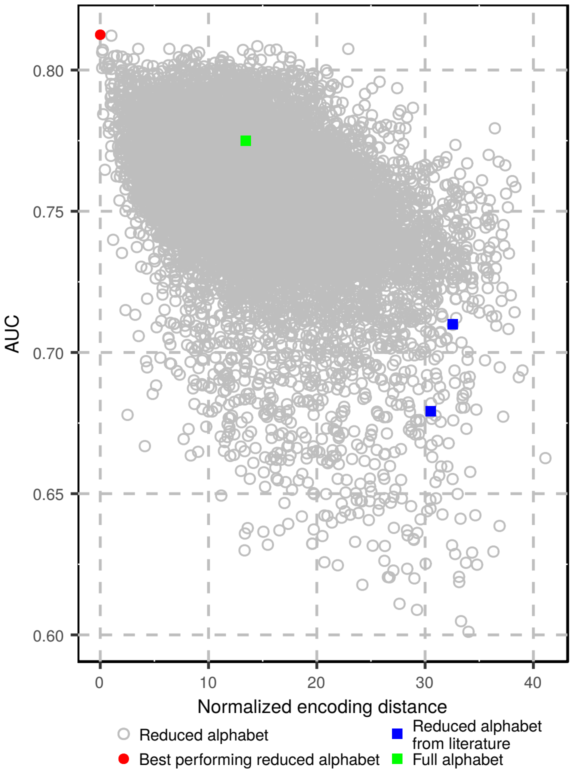
Considering only the AUC value, the most problematic testing data were sequences longer than 15 amino acids, where the classifiers trained on the shortest sequences available (six residues) were performing the best. That might indicate that our n-gram approach extracts the important features the better, the shorter are sequences in training data set. For example, in the amyloidogenic peptide of length 15, 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 are marked as amyloids. Despite this problem, the overall prediction of classifiers learned on the long sequences was still adequate, with the median values of AUC higher than 0.7 for every testing set.

In addition to the high AUC, the best encoding had also very good sensitivity and specificity regardless of the length of sequences in training and testing set (Fig. 3). The classifiers trained on the peptides of length 6 tend to have the best specificity, while predictors learned on the long sequences have the best sensitivity. Albeit the classifiers trained on six-residue-long sequences have generally better AUC, training on the sequences ranging from six to ten residues seem to yield the most balanced classifiers with optimal sensitivity and specificity.

We also evaluated classifiers based on the full amino acid alphabet. In most cases they were also placed in the fourth quartile considering their AUC value (Fig. 2). Nevertheless, they never predicted amyloidogenicity better than the selected encoding. It suggests that despite the predictive power of n-grams based on full alphabet, the encoding allows better generalization of the prediction rules and in the consequence a better performance.

Similarly to the best-performing encoding, the sensitivity of classifiers based on the full amino acid alphabet decreased with the length of sequences in the training data set (Fig. 3). Furthermore, these classifiers always seemed to have one of the worst sensitivities among all analyzed predictors, especially when tested on longer amyloids. It means that using the full amino acid alphabet it was easier to point the non-amyloidogenic sequences instead of recognizing amyloids.

The encodings found in the literature perform substantively worse than other analyzed reduced amino acid alphabets in all categories. It indicates that classical divisions of amino acids do not create groups suitable for the recognition of amyloids. This observation is well-supported by the specificity-sensitivity plot (Fig. 3), where classifiers trained with this encodings of amino acids groups with the worst performers.



figure

Figure 4. The encoding distance and AUC of reduced alphabets studied in the cross-validation. Classifiers with the smallest encoding distance to the best classifier have the highest AUC.

The best-performing encoding and important n-grams

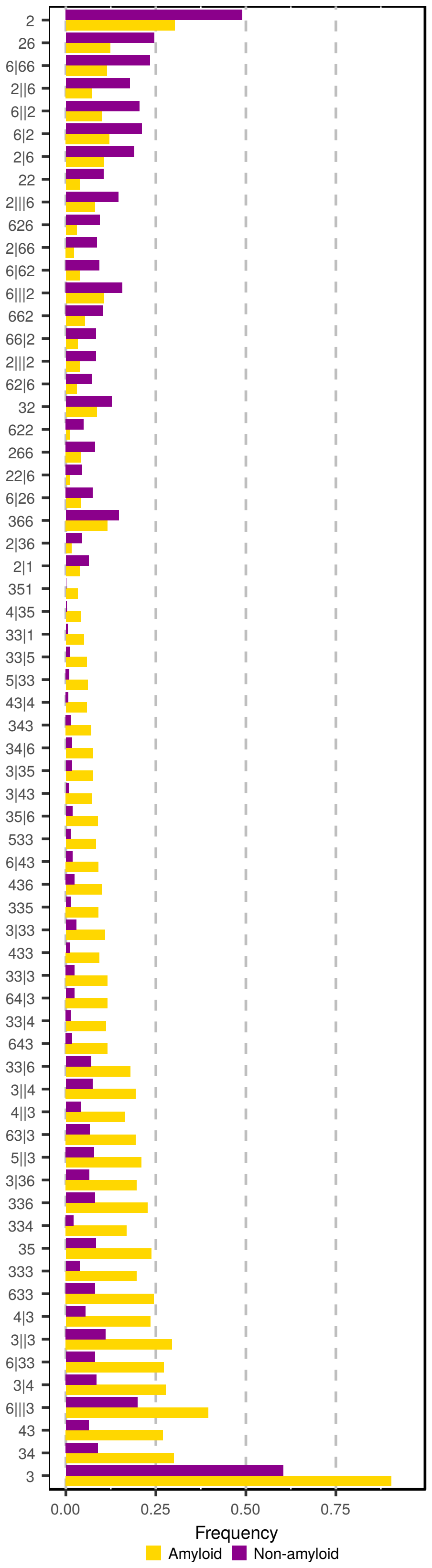
table

Table 3. The best-performing encoding.

Subgroup 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

The reduced amino acid alphabet chosen during the analysis has six subgroups (Tab. 3). Two subgroups, 3 and 4 contain strongly hydrophobic amino acids, while 1 and 5 are mostly hydrophilic. In addition to this, subgroup 1 and 4 have amino acid with the highest average flexibility. The least flexible amino acids are in subgroup 5. The most polarizable amino acids belong to subgroup 4, while the lowest polarizability is typical for glycine, the only amino acid in subgroup 1. The glycine has also the highest thermodynamic beta sheet propensity, which has the lowest values for subgroups 3 and 4.

In total, eleven combinations of physicochemical properties created the best performing reduced amino acid alphabet. Only four features appeared in all combinations: hydrophobicity index (Argos et al., 1982), average flexibility indices (Bhaskaran and Ponnuswamy, 1988), polarizability



figure

Figure 5. 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 (see Tab. 3). A vertical bar represents a gap in a n-gram.

parameter (Charton and Charton, 1982) and thermodynamic β-sheet propensity (Kim and Berg, 1993).

We calculated encoding distances between the best-performing reduced amino acid alphabet and other encodings (Fig. 4). To compute the scale factor, we used described above features common for all encodings identical to the best-performing one. The value of AUC is significantly lower for more distant encodings (-0.4370 Pearson's correlation coefficient, p-value smaller than 2.2×10^{-16}). Such relationship between AUC and encoding distance confirms that only reduced alphabets sufficiently similar to the best-performing encoding are able to remove unnecessary diversity while preserving information important for recognition of amyloids.

We selected 65 n-grams that have p-values smaller than 0.05 in the all folds in all repetitions of cross-validation regardless of the length of sequences in the training set (see Fig. 5). The frequency of the n-grams was computed for all sequences derived from AmyLoad database. The n-grams typical of amyloidogenic sequences (with the highest frequency in amyloids) incorporate mostly highly hydrophobic, typical for β-structures amino acids from subgroups 3 and 4. The important n-grams occurring frequent in amyloids often have repeats of 3, suggesting that the presence of amino acids belonging to this subgroup might be one of the most effective predictors of amyloidogenicity.

N-grams typical of non-amyloidogenic peptides have mostly elements belonging to subgroups 2 and 6. The amino acids of these subgroups are strongly hydrophilic and highly flexible, in the opposite to the residues typical for amyloids.

Benchmark of AmyloGram

table

Table 4. Results of benchmark on *pep424* data set for AmyloGram, PASTA2, FoldAmyloid and random forest predictor learned on n-grams extracted without any amino acid encoding from the sequences of the length specified in the brackets (FA).

Classifier	AUC	MCC	Sensitivity	Specificity
AmyloGram	**0.8972**	**0.6307**	**0.8658**	0.7889
FA (6)	0.8411	0.5427	0.4966	**0.9593**
FA (6-10)	0.8581	0.5698	0.7517	0.8259
FA (6-15)	0.8610	0.5490	0.8188	0.7519
PASTA2	0.8550	0.5227	0.7987	0.7444
FoldAmyloid	0.7351	0.4526	0.7517	0.7185

The benchmark covered Amylogram as well as two best-performing peer-reviewed predictors of amyloidogenicity: PASTA2 (Walsh et al., 2014) and FoldAmyloid (Garbuzynskiy et al., 2010). We analyzed Area Under the Curve (AUC), Matthew's Correlation Coefficient (MCC), Sensitivity and Specificity (see Tab. 4). We used default settings for FoldAmyloid, while PASTA2 evaluated input data in the 'Peptides' mode.

In case of this dataset, n-gram extraction method was efficient enough to produce classifiers able to outperformed published methods. Two of three n-gram based classifiers trained on the full alphabet have AUC higher than PASTA2 and all three were more successful than FoldAmyloid. They also maintained they high Specificity as seen previously during cross-validation.

Although the proposed n-gram extraction creates accurate classifiers, the encoding of amino acids further increases the efficiency of prediction. AmyloGram has the highest AUC, MCC and Sensitivity among all tested classifiers. Is has lower specificity than two classifiers trained on the full alphabet, but still outperforms other published method in this category. It is important to highlight that AmyloGram is the most of analyzed classifiers, having the best Specificity/Sensitivity trade-off, as indicated by the value of MCC.

CONCLUSION

Thanks to the reduction of amino acid alphabet, we were able to create the efficient predictor of amyloidogenic sequences called AmyLoad. One of the strength of our approach is its highly interpretable outcome, which hopefully sheds new light on the process of amyloid aggregation.

The idea of using the reduced amino acid alphabets is not new, but we employed innovative framework to generate and validate several thousands of possible amino acid encodings. Due to this approach, we are able to specify important physicochemical properties that define the best-performing alphabet. We confirmed the relevance of properties commonly associated with amyloidogenicity as hydrophobicity and discover new ones, as flexibility.

Our analysis was completed with the extraction of important n-grams, which might be interpreted as short motifs highly relevant to amyloidogenicity or nonamyloidogenicity of the peptide. 65 important n-grams revealed that mostly alifatic and nonpolar amino acids as isoleucine, leucine and valine are responsible for the hydrophobic character of amyloids. Only in their neighborhood, the presence of aromatic and hydrophobic amino acids (phenylalanine, tyrosine, tryptophan) is also a sign of a potential amyloid.

Since the best-performing classifier was trained on the alphabet of length six, but still outperformed predictors learning on the raw amino acid sequence, we cannot surely determine the optimal length of a reduced amino acid alphabet for detection of amyloids. It is plausible that such alphabet is longer than six residues. Nevertheless, the alphabet of length six is enough to find n-grams separating efficiently amyloids and non-amyloids.

FUNDING

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 partially funded by the KNOW Consortium and National Science Center (2015/17/N/NZ2/01845).

REFERENCES

Sergiy O. Garbuzynskiy, Michail Yu Lobanov, and Oxana V. Galzitskaya. FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics (Oxford, England)*, 26(3):326–332, February 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp691.

Allen W. Bryan, Charles W. O'Donnell, Matthew Menke, Lenore J. Cowen, Susan Lindquist, and Bonnie Berger. STITCHER: Dynamic assembly of likely amyloid and prion -structures from secondary structure predictions. *Proteins*, 80(2):410–420, February 2012. ISSN 1097-0134. doi: 10.1002/prot.23203.

Lukasz Goldschmidt, Poh K. Teng, Roland Riek, and David Eisenberg. Identifying the amylome, proteins capable of forming amyloid-like fibrils. *Proceedings of the National Academy of Sciences*, 107(8):3487–3492, February 2010. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0915166107. URL http://www.pnas.org/content/107/8/3487.

Charles W. O'Donnell, Jrme Waldisphl, Mieszko Lis, Randal Halfmann, Srinivas Devadas, Susan Lindquist, and Bonnie Berger. A method for probing the mutational landscape of amyloid structure. *Bioinformatics*, 27(13):i34–i42, July 2011. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btr238. URL http://bioinformatics.oxfordjournals.org/content/27/13/i34.

Jacinte Beerten, Joost Van Durme, Rodrigo Gallardo, Emidio Capriotti, Louise Serpell, Frederic Rousseau, and Joost Schymkowitz. WALTZ-DB: a benchmark database of amyloidogenic hexapeptides. *Bioinformatics (Oxford, England)*, 31(10):1698–1700, May 2015. ISSN 1367-4811. doi: 10.1093/bioinformatics/btv027.

Pawel Gasior and Malgorzata Kotulska. FISH Amyloid a new method for finding amyloidogenic segments in proteins based on site specific co-occurence of aminoacids. *BMC Bioinformatics*, 15(1):54, February 2014. ISSN 1471-2105. doi: 10.1186/1471-2105-15-54. URL http://www.biomedcentral.com/1471-2105/15/54/abstract.

Jerzy Stanislawski, Malgorzata Kotulska, and Olgierd Unold. Machine learning methods can replace 3d profile method in classification of amyloidogenic hexapeptides. *BMC Bioinformatics*, 14:21, 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-21. URL http://dx.doi.org/10.1186/1471-2105-14-21.

Rafael Zambrano, Michal Jamroz, Agata Szczasiuk, Jordi Pujols, Sebastian Kmiecik, and Salvador Ventura. AGGRESCAN3d (A3d): server for prediction of aggregation properties of protein structures. *Nucleic Acids Research*, page gkv359, April 2015. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkv359. URL http://nar.oxfordjournals.org/content/early/2015/04/16/nar.gkv359.

Lynne Reed Murphy, Anders Wallqvist, and Ronald M. Levy. Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Engineering*, 13(3):149–152, March 2000. ISSN 1741-0126, 1741-0134. doi: 10.1093/protein/13.3.149. URL http://peds.oxfordjournals.org/content/13/3/149.

Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001. ISSN 0885-6125, 1573-0565. doi: 10.1023/A:1010933404324. URL

Pawel P. Wozniak and Malgorzata Kotulska. AmyLoad: website dedicated to amyloidogenic protein fragments. *Bioinformatics (Oxford, England)*, June 2015. ISSN 1367-4811. doi: 10.1093/bioinformatics/btv375.

Joe H. Ward Jr. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244, March 1963. ISSN 0162-1459. doi: 10.1080/01621459.1963.10500845. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500845.

Shuichi Kawashima, Piotr Pokarowski, Maria Pokarowska, Andrzej Kolinski, Toshiaki Katayama, and Minoru Kanehisa. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Research*, 36(suppl 1):D202–D205, January 2008. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkm998. URL http://nar.oxfordjournals.org/content/36/suppl$_1/D$202.

R. Bhaskaran and P.k. Ponnuswamy. Positional flexibilities of amino acid residues in globular proteins. *International Journal of Peptide and Protein Research*, 32(4):241–255, October 1988. ISSN 1399-3011. doi: 10.1111/j.1399-3011.1988.tb01258.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1399-3011.1988.tb01258.x/abstract.

K. Nishikawa and T. Ooi. Radial locations of amino acid residues in a globular protein: correlation with the sequence. *Journal of Biochemistry*, 100(4):1043–1047, October 1986. ISSN 0021-924X.

Anna Radzicka and Richard Wolfenden. 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, March 1988. ISSN 0006-2960. doi: 10.1021/bi00405a042. URL http://dx.doi.org/10.1021/bi00405a042.

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

1097-0134. doi: 10.1002/prot.10584.

Pawel P. Wozniak and Malgorzata Kotulska. Characteristics of protein residue-residue contacts and their application in contact prediction. *Journal of Molecular Modeling*, 20(11), 2014. ISSN 1610-2940. doi: 10.1007/s00894-014-2497-9. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4221654/.

Minoru I. Kanehisa and Tian Yow Tsong. Local hydrophobicity stabilizes secondary structures in proteins. *Biopolymers*, 19(9):1617–1628, September 1980. ISSN 1097-0282. doi: 10.1002/bip.1980.360190906. URL http://onlinelibrary.wiley.com/doi/10.1002/bip.1980.360190906/abstract.

M. Prabhakaran. The distribution of physical, chemical and conformational properties in signal and nascent peptides. *The Biochemical Journal*, 269(3): 691–696, August 1990. ISSN 0264-6021.

C. A. Kim and J. M. Berg. Thermodynamic beta-sheet propensities measured using a zinc-finger host peptide. *Nature*, 362(6417):267–270, March 1993. ISSN 0028-0836. doi: 10.1038/362267a0.

P. Argos, J. K. Rao, and P. A. Hargrave. Structural prediction of membrane-bound proteins. *European journal of biochemistry / FEBS*, 128(2-3):565–575, November 1982. ISSN 0014-2956.

R. M. Sweet and D. Eisenberg. Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *Journal of Molecular Biology*, 171(4):479–488, December 1983. ISSN 0022-2836.

Akinori Kidera, Yasuo Konishi, Masahito Oka, Tatsuo Ooi, and Harold A. Scheraga. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *Journal of Protein Chemistry*, 4(1):23–55, February 1985. ISSN 0277-8033, 1573-4943. doi: 10.1007/BF01025492. URL http://link.springer.com/article/10.1007/BF01025492.

S. D. Black and D. R. Mould. Development of hydrophobicity parameters to analyze proteins which bear post- or cotranslational modifications. *Analytical Biochemistry*, 193(1):72–82, February 1991. ISSN 0003-2697.

Marvin Charton and Barbara I. Charton. The structural dependence of amino acid hydrophobicity parameters. *Journal of Theoretical Biology*, 99(4):629–644, December 1982. ISSN 0022-5193. doi: 10.1016/0022-5193(82)90191-6. URL http://www.sciencedirect.com/science/article/pii/0022519382901916.

Joan Pontius, Jean Richelle, and Shoshana J. Wodak. Deviations from Standard Atomic Volumes as a Quality Measure for Protein Crystal Structures. *Journal of Molecular Biology*, 264(1):121–136, November 1996. ISSN 0022-2836. doi: 10.1006/jmbi.1996.0628. URL http://www.sciencedirect.com/science/article/pii/S0022283696906282.

K. Takano and K. Yutani. 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, August 2001. ISSN 0269-2139.

Jian Tian, Ningfeng Wu, Jun Guo, and Yunliu Fan. Prediction of amyloid fibril-forming segments based on a support vector machine. *BMC Bioinformatics*, 10(1):1–8, 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-S1-S45. URL http://dx.doi.org/10.1186/1471-2105-10-S1-S45.

Malgorzata Kotulska and Olgierd Unold. On the amyloid datasets used for training PAFIG how (not) to extend the experimental dataset of hexapeptides. *BMC Bioinformatics*, 14:351, December 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-351. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3879009/.

Marvin N. Wright and Andreas Ziegler. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *arXiv:1508.04409 [stat]*, August 2015. URL http://arxiv.org/abs/1508.04409. arXiv: 1508.04409.

Erich L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer New York, August 2008. ISBN 978-0-387-98864-1.

Ian Walsh, Flavio Seno, Silvio C. E. Tosatto, and Antonio Trovato. PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Research*, page gku399, May 2014. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gku399. URL http://nar.oxfordjournals.org/content/early/2014/05/21/nar.gku399.