PACT - Prediction of Amyloid Cross-interaction by Threading

Jakub W. Wojciechowski[1,*], Witold Szczurek[1], and Małgorzta Kotulska[1,*]

[1]Department of Biomedical Engineering, Faculty of Fundamental Problems of Technology, Wrocław University of Science and Technology, 50-370 Wrocław, Poland
[*]jakub.wojciechowski@pwr.edu.pl, malgorzata.kotulska@pwr.edu.pl

ABSTRACT

Amyloids are protein aggregates usually associated with their contribution to several diseases e.g., Alzheimer's and Parkinson's. However, they are also beneficially utilized by many organisms in physiological roles, such as microbial biofilm formation or hormone storage. Recent studies showed that an amyloid aggregate can affect aggregation of another protein. Such cross-interactions may be crucial for understanding comorbidity of amyloid diseases or influence of microbial amyloids on human amyloidogenic proteins. However, due to demanding experiments, understanding of interaction phenomenon is still limited. Moreover, no dedicated computational method to predict potential amyloid interactions has been available until now. Here, we present PACT - a computational method for prediction of amyloid cross-interactions. The method is based on modeling a heterogenous fibril formed by two amyloidogenic peptides. Stability of the resulting structure is assessed using a statistical potential that approximates energetic stability of a model. Importantly, the method can work with long protein fragments and, as a purely physicochemical approach, it relies very little on training data. PACT was evaluated on data collected in AmyloGraph database and it achieved high values of AUC (0.88) and F1 (0.82). The new method opens a possibility of high throughput studies of amyloid interactions. We used PACT to study interactions of CsgA, a bacterial biofilm protein from several bacterial species inhabiting human intestines, and human Alpha-synuclein protein which is involved in the onset of Parkinson's disease. We show that the method correctly predicted the interactions and highlighted the importance of specific regions in both proteins.
The tool is available as a web server at: https://pact.e-science.pl/pact/. The local version can be downloaded from: https://github.com/KubaWojciechowski/PACT

Introduction

Pathological misfolding and aggregation of proteins is a hallmark of a number of devastating disorders, including major public health challenges like Alzheimer's and Parkinson's diseases[1], type II diabetes[2], as well as some cancers[3]. These diseases not only share a similar molecular mechanism, but they also often co-occur in the same patients. Among others, comorbidities were observed between Alzheimer's disease and type II diabetes[4,5] and Alzheimer's and Parkinson's diseases[6]. One of the possible explanations of this phenomenon could be related to amyloid cross-interactions. Amyloids are insoluble protein aggregates characterized by an exceptional stability due to the tight packing of monomers, resulting in characteristic pattern in X-ray diffraction experiments.[7]. Despite significant structural similarities shared by all amyloids, their sequences are surprisingly diverse and have little homology[8]. More recent studies have shown that in some cases presence of amyloid aggregates can affect the aggregation rate of other proteins[9]. Later, it was observed that interacting proteins can form heterogeneous fibers consisting of molecules of both interaction partners. Such cross-interactions were then identified between numerous proteins, including those involved in type II diabetes and neurodegenerative diseases. This shed new light on potentially new aspects regarding the origin of comorbidity of these disorders[10]. Despite the importance of this process, its mechanisms are still poorly understood, which can be attributed to a limited number of experimental data. As a result, interactions of only a few well-described proteins, such as Amyloid-beta (Abeta), islet amyloid polypeptide, or Alpha-synuclein, have been very extensively studied and they contributed to a majority of data. This may introduce a bias in available data. Despite the difficulties, it was shown that proteins with similar sequences are more likely to interact, however, many counterexamples were also shown[10]. The studies highlight the importance of the structural compatibility of amyloid cores.

The main limitation regarding experimental studies of amyloid aggregation and their interactions is that they require expensive and time consuming methods. In practice, biochemical assays based on binding of Congo Red[11], Thioflavin T[12] and infrared spectroscopy are frequently used. Especially the last method is widely applied, due to its simplicity and efficiency[13]. Another approach involves direct observation of fibers using high resolution imaging techniques, such as electron microscopy[14]

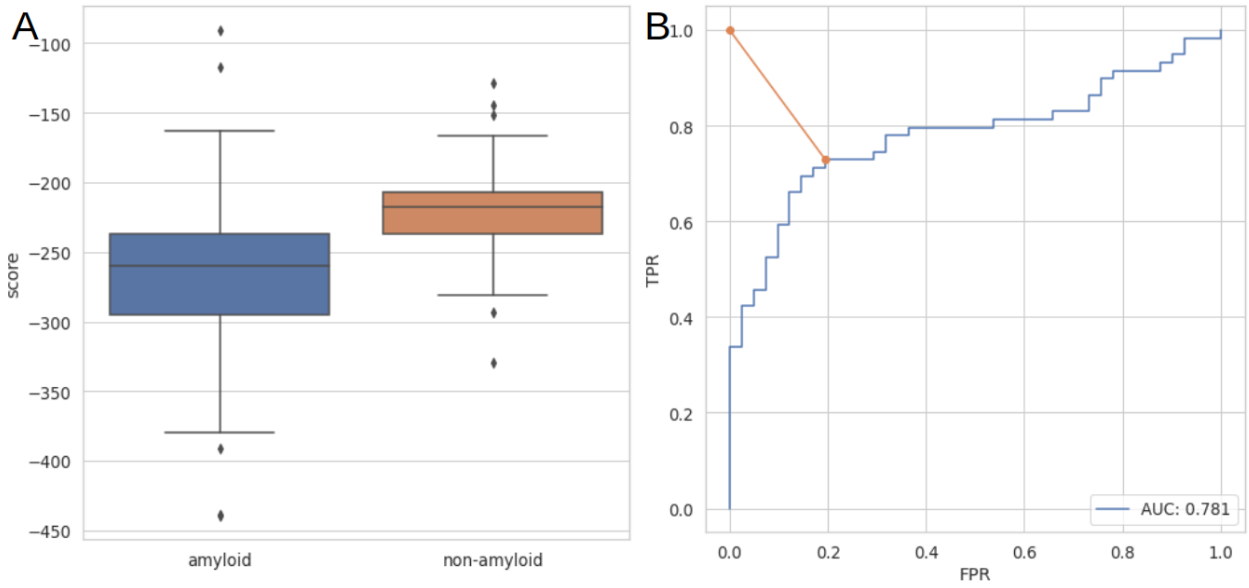

Figure 1. A) Distribution of *ndope* score for models of amyloidogenic and non-amyloidogenic peptides. B) ROC curve for amyloid vs non-amyloid classification. The orange line represents the distance between perfect classification point (0,1) and the chosen threshold

and atomic force microscopy[15]. Since different methods rely on different approaches, their results might differ in some cases. More importantly, all of them are expensive and time consuming. Experiments are hampered by difficulties in handling amyloids, including their low solubility, rapid aggregation, and need for their high-purity[16]. Currently, the use of experimental methods for identification of all amyloids in genome wide studies would be impossible. To address this problem several computational methods have been proposed based on different approaches (reviewed in[17] and[18]), starting from structural modeling[19], statistical analysis of the sequence including FoldAmyloid[20] and FishAmyloid[21], physicochemical models like PASTA 2.0[22], machine learning techniques such as APPNN[23] and AmyloGram[24]. Furthermore, there are methods combining both approaches such as PATH[25] and Cordax[26]. Finally, some methods, like Aggrescan 3D[27], utilize information about protein structure. It was also shown that bioinformatics techniques are quite robust and capable of identification even some misannotated data despite being trained on them[28]. Unfortunately, none of these methods can predict amyloid interactions.

Here, we present a new computational method PACT (Prediction of Amyloid Cross-interaction by Threading) designed for identification of potentially interacting amyloid pairs. The method is based on the molecular threading applied to the potential complex of interacting amyloids and assessment of the stability of obtained molecular model.

Results

The main assumption of the method is that interactions between amyloidogenic fragments that cross-interact, threaded into an amyloid fiber structure, would result in a heterogeneous aggregate which is more stable, thus energetically more favorable than a non-interacting pair. In PACT, we use the Modeller[29] software for threading a query sequence on the structure of amyloid fiber formed by Islet Amyloid Polypeptide (IAPP)[30]. To asses obtained models we have proposed *ndope* score, which is a normalized version of DOPE statistical potential implemented in the Modeller software.

PACT correctly identifies amyloid prone peptides

In the first step, we focused on prediction of homoaggregation, which can be considered a special case of heteroaggregation. We compared *ndope* scores obtained for models of potential homoagregates of amyloidogenic and non-amyloidogenic peptides, for which the sequences were obtained from the AmyLoad database[31]. The majority of models obtained for amyloidogenic peptides showed much lower scores (meaning more stable structures) in comparison with non-amyloids and their first quartiles of the scores were well separated (Fig.1). Differences between both groups were statiscally significant. Based on the Mann–Whitney U test we were able to reject the hypothesis that the distributions of both populations are identical ($p = 2.48e-8$). Considering energy difference, we built a threshold based classifier. The classification threshold was chosen based on the Receiver Operating Characteristic (ROC) curve, as a point on the curve closest to the point (0,1), representing perfect classification (Fig. 1B). The chosen value of *ndope* score was -242. If used merely for distingushing amyloids from non-amyloids, such a classifier was

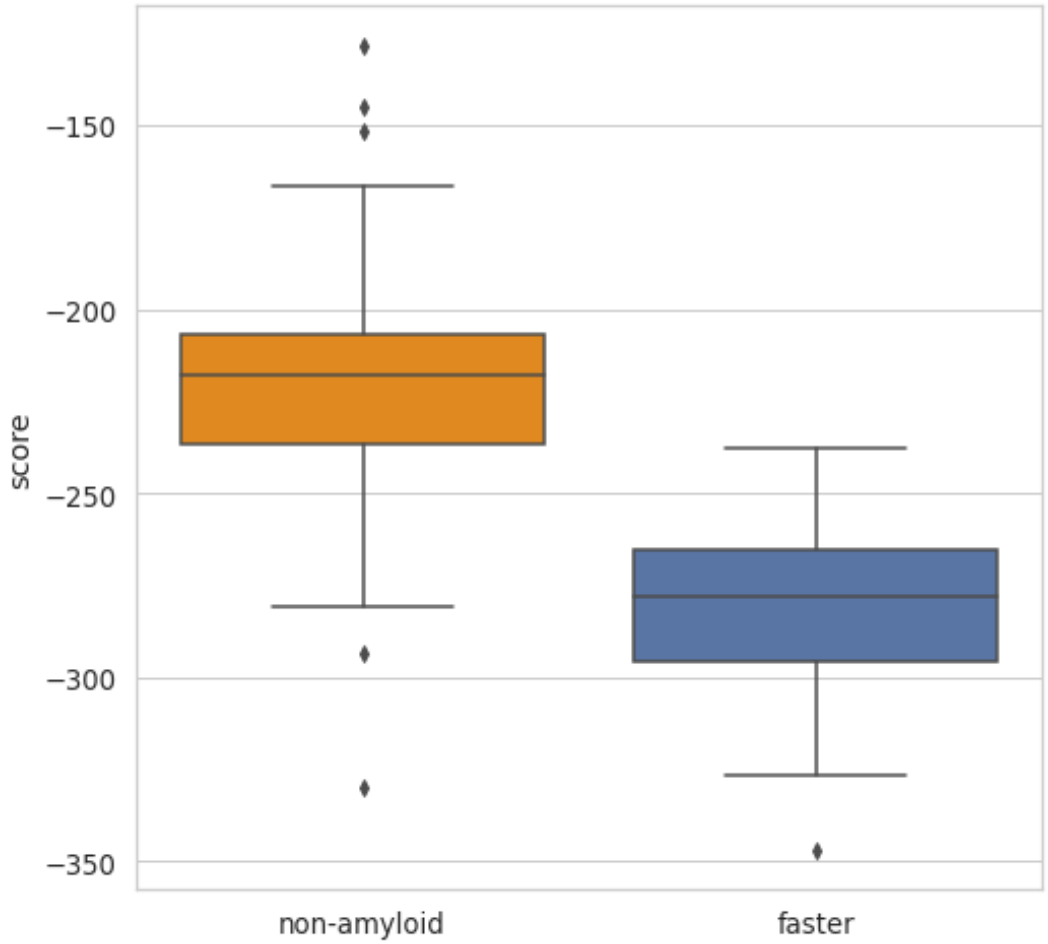

Figure 2. The score *ndope* for models of interacting identical non-amyloidogenic peptides (negative set) and interacting pairs resulting in increased aggregation rates (*faster* set).

able to achieve Area Under ROC Curve (AUC) of 0.73 and *Accuracy* of 0.77. Moreover, high values of *Sensitivity* (0.73) and *Specifictity* (0.86) were obtained. Such results are comparable with state-of-the-art amyloid predictors on the same data set (Sup Tab S1).

We also tested if the method is capable of recognizing amyloid propensity in functional amyloids, which pose a major problem for most predictors due to their under-representation in databases of amyloids. We tested the performance of the method on imperfect repeats of CsgA protein from *Escherichia coli* and *Salmonella enterica*[13] (Sup Fig S1). Aggregation prone regions of this protein (R1, R3, and R5) scored much lower than non-amyloidogenic regions (R2 and R4) from *Escherichia coli*. On this data, PACT achieved the accuracy of 0.9. Furthermore, the observed difference between *ndope* score for R4 fragments from *Escherichia coli* and *Salmonella enterica* corresponds very well to the difference in their aggregation propensities[13].

The results showed that the method can accurately predict aggregation prone peptides of varying lengths. Furthermore, it can be utilized to detect functional amyloids.

PACT predicts amyloid cross-interactions

We used similar methodology to predict cross-interactions of amyloid peptides, which is the main purpose of the method. The *ndope* scores of heteroaggregates consisting of pairs of peptides whose cross-interactions resulted in faster aggregation were compared with non-amyloidogenic pairs of peptides (Fig.2). A similar analysis was performed for pairs of peptides whose cross-interactions resulted in slower aggregation (Sup Fig S3). In both cases, models of heterologous aggregates resulting from cross-interactions showed lower values of *ndope* scores than non-amyloids, and well separated first quartiles of their scores (Fig.2). Furthermore, in both cases differences between groups were statistically significant (Mann–Whitney U test, $p = 5.54e-16$ for *faster* vs *negative* and $p = 2.19e-15$ *slower* vs *negative* cases. Therefore, we built the threshold based classifier using the approach described in the previous section.

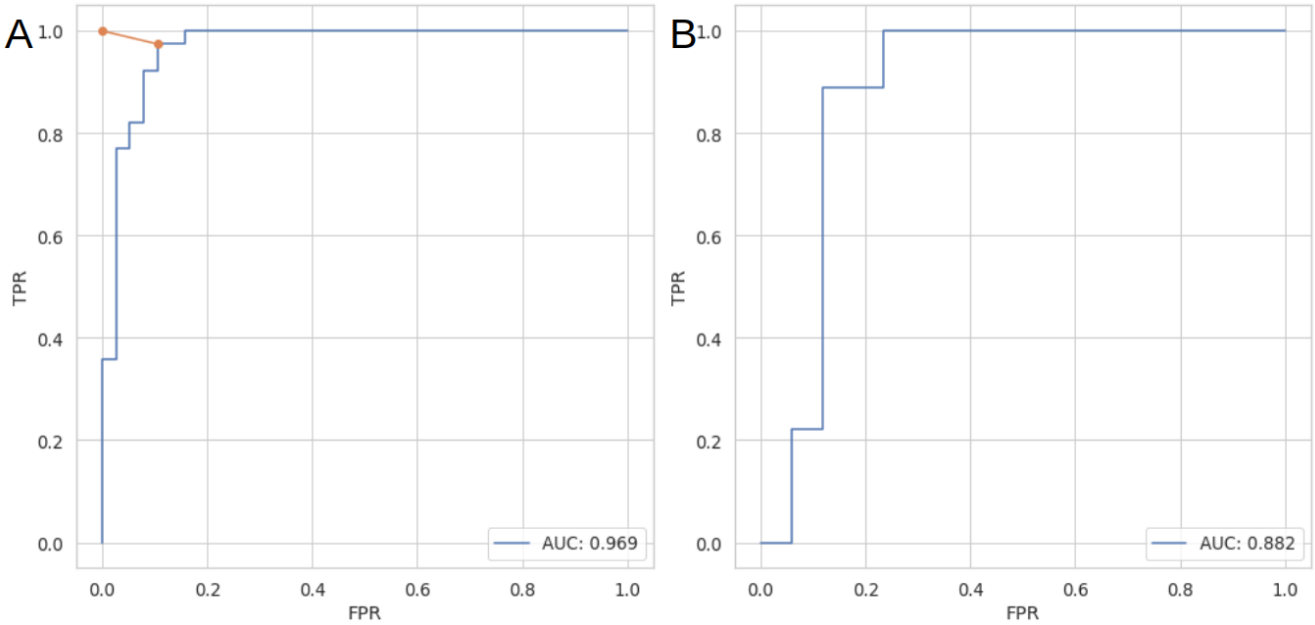

Figure 3. ROC curves for classification of non-aggregating and cross interacting pairs resulting in faster aggregation on (A) training and (B) test set.

Table 1. Performance of PACT on cross validation and independent test set for classification of non-aggregating and cross interacting pairs resulting in faster aggregation

	Acc [std]	Sens [std]	Spec [std]	F1 [std]	MCC [std]
Cross-validation	0.90 [0.05]	0.91 [0.03]	0.90 [0.08]	0.90 [0.04]	0.80 [0.06]
Test set	0.83	0.78	0.88	0.82	0.66

To assess the performance and choose the optimal threshold value, ROC curves were calculated for both cases; *faster* rate vs *negative* (Fig.3) and *slower* vs *negative* (Sup Fig S3) on both training and test sets. To minimize the impact of the data choice, we performed k-folds cross-validation with k=5 on the training set and calculated several metrics describing the performance of the method (Tab. 1). The same metrics were then calculated on an independent test set. The same analysis was performed for the case of prediction of interactions resulting in slower aggregation (Sup Tab S2). Chosen *ndope* thresholds were very similar in both scenarios, namely -256 and -245 for faster vs negative and slower vs negative respectively. PACT performed well on both cross-validation and independent test set. It achieved the *Accuracy* of 0.83 and 0.80 on test sets of *faster* vs *negative* and *slower* vs *negative* cases, respectively. In all cases, the results obtained on the test set were within the value of one standard deviation range from the mean values obtained with the cross-validation procedures. The method performance was quite similar in both *faster* vs *negative* and *slower* vs *negative* scenarios. However, due to the smaller data set size, larger standard deviation was obtained for *slower* vs *negative* scenario (Table 1). The results show that the method can predict whether two peptides can cross-interact but cannot distinguish between types of interactions with regard to their rate.

PACT is robust to bias in data

A serious problem with data regarding interacting amyloids, available in the literature and, therefore, in our dataset is large overrepresentation of interactions concerning Abeta peptide. This may cause overfiting of the method to this group of sequences. To assess its effect, we analyzed scores for interactions between different Abeta variants (Sup Fig S4). Observed scores for Abeta pairs are within a range of values observed for the remaining pairs, and therefore they should not have a significant effect on the performance of the method. Those pairs showed relatively narrow distribution of *ndope* values, centered slightly below *ndope* value of -275, which is relatively close to identified classification threshold of -256, while the remaining interacting pairs showed even lower scores.

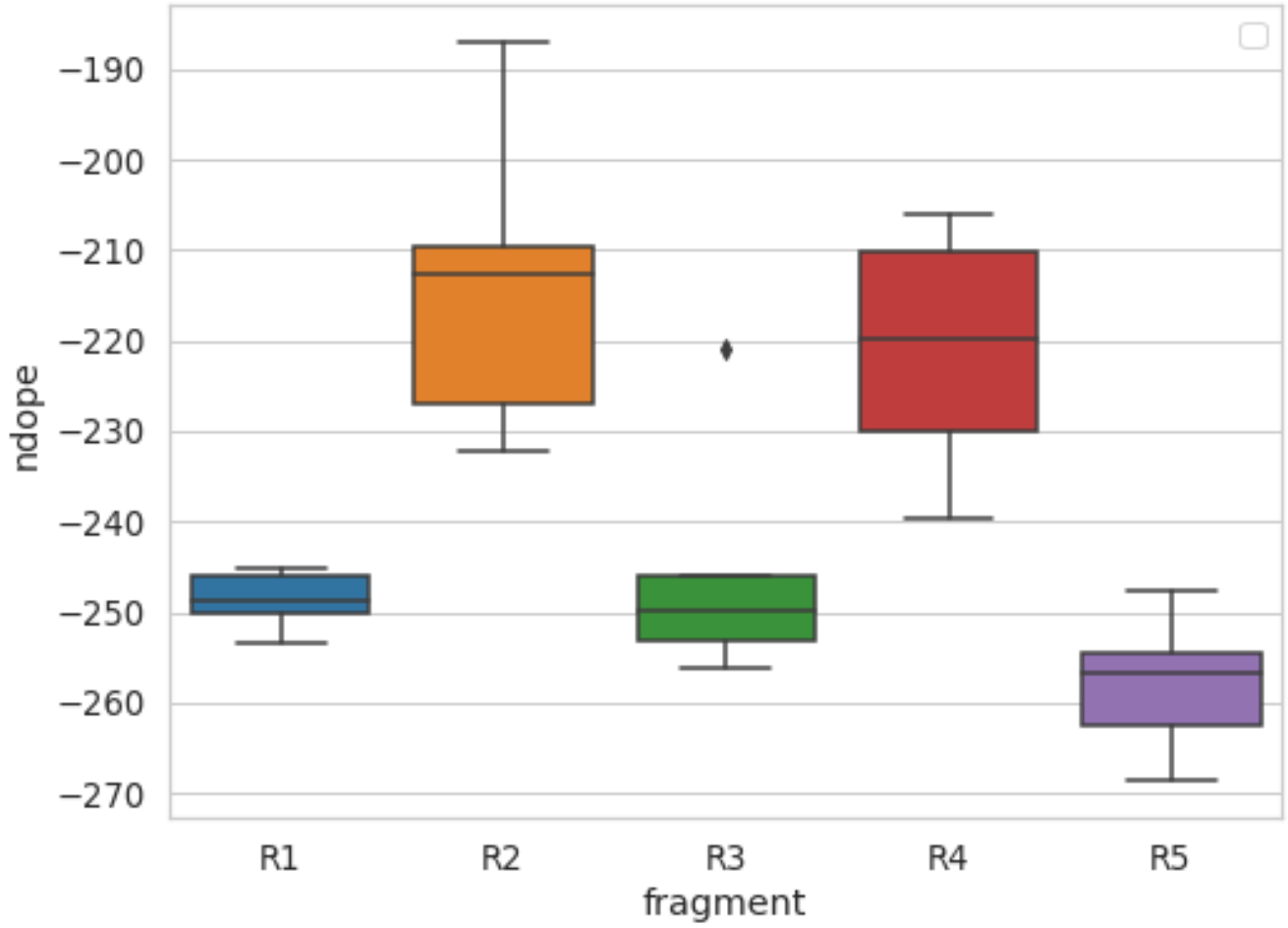

Figure 4. Lowest *ndope* scores for interactions of R1-R5 repeats from each of CsgA proteins with Alpha-synuclein.

Interactions between bacterial amyloids and Alpha-synuclein

In recent years plenty of studies highlighted the connection between gut microbiome composition and the onset of many diseases including neurodegenerative ones like Alzheimer's and Parkinson's diseases[32]. Despite extensive studies, understanding of molecular mechanisms of this connection remains elusive. One of the possible explanations for this relates functional amyloids from bacteria and human disease related amyloids through the cross-interaction theory. The aggregation of the bacterial ones could speed up the aggregation of the disease related proteins, leading to the disorder[33]. This hypothesis seems consistent with results obtained by Chen and co-workers, who discovered increased production and aggregation of Alpha-synuclein in rats exposed to bacterial strains producing biofilm related functional amyloids[34].

In order to better understand this connection we studied possible interactions between bacterial functional amyloid CsgA and human Alpha-synuclein, whose aggregation is a hallmark of Parkinson's disease. We used PACT to predict interactions of CsgA protein from five different organisms found in human microbiome; *Escherichia coli* (EC), *Hafnia alvei* (HA), *Yokenella regensburgei* (YR), *Citrobacter youngae* (CY), and C*edecea davisae* (CD) with human Alpha-synuclein which were recently studied experimentally by Bhoite and coworkers[35].

The sequence of Alpha-synuclein was divided into overlapping fragments of length 20 amino acids and their interactions with R1-R5 repeats of each of CsgA proteins were tested. Consistently with experimental results, all of the studied CsgA variants were predicted to interact with Alpha-synuclein. Among CsgA proteins' fragments, R1, R3 and R5 were predicted to interact, with R5 showing the best scores (Fig4). These results are consistent with our current state of knowledge about CsgA as the most aggregation prone regions in these proteins are R1, R3 ad R5. Furthermore, R5 fragment which showed the lowest *ndope* scores, is located at the protein surface, therefore it can interact without a need for major conformational changes. On the Alpha-synuclein part the best scoring region was located between positions 32 and 56 (Sup Fig S5). This region was recently shown to be of crucial importance for the aggregation of the protein[36].

Code Availability

PACT was implemented as an open source Python module, available at GitHub repository: https://github.com/KubaWojciechowski/PACT. For users' convenience we prepared a docker container for the application, as well as the web server: https://pact.e-science.pl/pact/

Discussion

High throughput identification of amyloid cross-interaction is an important step toward our understanding of this mechanism. It can allow for a better understanding of the principles governing the process but it can also be used to identify novel cases of amyloid interactions. Such capabilities can shed light on a possible mechanism responsible for the comorbidity of devastating disorders.

We proposed the first computational method for predicting amyloid cross-interactions. It is based on a highly interpretable and well-established physicochemical model which is not heavily dependent on training data. This feature is especially important since the available data contains a strong interest bias toward interactions of a few popular amyloids related to neurodegenerative diseases, such as Abeta. We have carefully studied the effect of this overrepresentation and showed that it does not affect the performance of the method. Furthermore, good performance on functional amyloids suggests that the method is very robust and can be effectively used on a wide range of sequences. In total, PACT achieved a high accuracy of 0.83 and 0.80 on the independent test sets of interactions increasing and decreasing aggregation rates. On both sets the method achieved high AUC values of 0.88 and 0.89, and F1 values of 0.82 and 0.77. On the other hand, since both cases were characterized by similar interaction energies, the method cannot distinguish between enhancement and inhibition of aggregation cases. These results suggest that both processes may be driven by similar mechanisms. This issue was addressed in a recently published work by Louros and coworkers[37], who applied a somewhat similar approach to study the effect of point mutations on aggregation characteristics.

We used PACT to predict interactions of bacterial functional amyloid CsgA from different species with human Alpha-synuclein. Not only our results are in a good agreement with recently published data but they can also indicate which regions can drive the cross-interactions between both proteins. Identification of potentially interacting regions can provide an important insight into possible mechanism of the process and guide the future experiments.

Apart from identification of amyloid cross-interactions, the proposed method is also capable of reliably predicting amyloid prone regions in proteins with comparable accuracy to the state-of-the-art techniques. Furthermore, it overcomes their major limitations regarding the identification of functional amyloids. Unlike most of the currently available amyloid predictors, it does not rely on scanning of a query sequence with a very short sliding window.

Methods

The main assumption of the method is that interactions between amyloidogenic fragments that cross-interact, threaded into an amyloid fiber structure, would result in a heterogeneous aggregate which is more stable, thus energetically more favorable than a non-interacting pair. A somewhat similar assumption was successfully applied in our previous work to predict the aggregation of short amyloidogenic fragments[25], although the current approach differs in other aspects of the method and the objectives. In PACT, we use the Modeller software for threading a query sequence on the structure of amyloid fiber formed by Islet the Amyloid Polypeptide (IAPP)[30]. To asses obtained models we have proposed *ndope* score, which is a normalized version of the DOPE statistical potential implemented in the Modeller software.

Data sets

To build and test the method we used the following datasets:

- the set of 90 amyloidogenic (*amyloid*) and 55 non-amyloidogenic (*non-amyloid*) peptides of lengths between 14 and 45 from the AmyLoad database[31].
- the set of 119 pairs of peptides, which enhance (*faster* dataset) and 73 which slow down (*slower* dataset) the aggregation of each other. Both from AmyloGraph database[38]. After removal of identical records 57 and 55 pairs of peptides which enhance and slow down the aggregation of each other respectively.

The first two sets (*amyloid* and *non-amyloid*) were used to test the method on cases of homoaggregation i.e., identifying amyloid prone peptides.

For the prediction of cross-interactions, we used *faster*, *slower* and non-amyloid sets. The use of the set of non-aggregating peptides as the negative set in the interaction study was caused by the lack of a sufficient number of negative examples of

non-interacting amyloid pairs. This is a common problem in studies of protein-protein interactions since negative results are rarely published, which often creates a strong bias in biological data[39].

Datasets used in this study are available at GitHub repository:
https://github.com/KubaWojciechowski/PACT

Modeling

A query pair of sequences was threaded on the structure of amyloid fiber formed by Islet Amyloid Polypeptide (IAPP)[30]. In order to allow the method to deal with sequences of varying lengths, sequences shorter than the sequence of a template use only the main part of the template structure. In such cases, a shorter sequence is aligned to the middle of the template sequence (Fig 5A). This choice can be justified considering that most of the currently known amyloid fragments, which are longer than a few amino acids, share a similar beta-sheet turn architecture, commonly known as the beta arch. This assumption was previously successfully applied by Ahmed and coworkers to build the ArchCandy method for amyloidogenic region prediction[40]. PACT allows sequences to be marginally longer than the template and, as a result, can be used to study cross-interactions between peptides of lengths between 14 and 45. For each of the tested pairs, 10 different models, consisting of two chains of each interacting peptide (Fig 5B) were built using Modeller 9.24 model-multichain.py procedure with default parameters[29]. Then, the model with the lowest DOPE value was chosen for further analysis. Since the dataset consisted of fragments of varying lengths, we proposed a normalized DOPE score (*ndope*) defined as follows:

$$ndope = \frac{DOPE}{L} \tag{1}$$

where L is an average length of sequences used to build a given model. Then, *ndope* scores were compared between amyloids and non-amyloids, as well as between pairs of amyloids interacting with non-amyloids.

To choose the *ndope* threshold for the classification, Receiver Operating Characteristics (ROC) curve was calculated by applying different score thresholds and recording False Positive Ratio (FPR) and True Positive Ratio (TPR). The threshold closest to the (0,1) point (representing perfect classification) was chosen.

We also tested a variant of the method which utilized three different structural templates (PDB: 2nnt, 2e8d), however it did not improve the accuracy of the method but significantly increased the computational time. Therefore, this approach was finally abandoned.

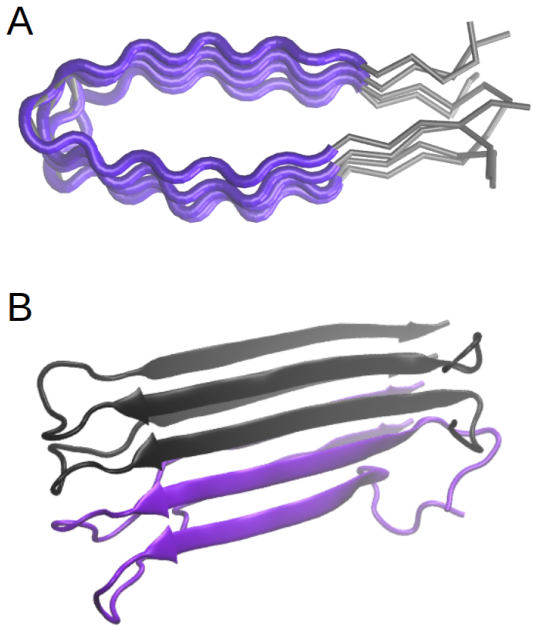

Figure 5. Schematic representation of the modeling procedure. A) In case when a query sequence is shorter than the template, only a part of it is used in modeling. B) The model of heterogenous fibril consists of two chains of each interacting peptide.

Assessment of performance and data analysis

All the data analysis was performed using Python 3.8 with Matplotlib[41], NumPy[42], Pandas[43], Scikit-Learn[44], and Seaborn[45] packages.

To test the performance of the proposed method, a dataset was randomly split into a training set and test set, which consisted of 30% of the data. Additionally, k-folds cross-validation (with k=5) was performed on the training data. Area Under ROC curve (AUC), *Accuracy* (ACC), *Sensitivity* (Sens), *Specificity* (Spec) and Matthew Correlation Coefficient (MCC) were used to assess the performance of the method

Effect of Amyloid-beta variants

For the analysis of the effect of over represented amyloid beta pairs, we divided the *faster* data set into two subsets: one containing only pairs were both interacting peptides were variants of amyloid beta (16 pairs) (*abeta*), and the set of remaining pairs (39 pairs) (*no abeta*).

Interactions between bacterial amyloids and Alpha-synuclein

Modeling the interactions between Alpha-synuclein and CsgA proteins was performed using human Alpha-synuclein sequence (Uniprot id: P37840) and CsgA protein from five different organisms found in human microbiome; *Escherichia coli* (EC) (Uniprot id: P28307), *Hafnia alvei* (HA) (Uniprot id: G9Y7N6) , *Yokenella regensburgei* (YR) (Uniprot id: A0A6H0K4L9), *Citrobacter youngae* (CY) (Uniprot id: A0A549VPM7), and C*edecea davisae* (CD) (Uniprot id: S3IYN9). The sequence of Alpha-synuclein was divided into overlaping subsequences of length 20. CsgA variants were split into five non-overlaping fragments R1-R5 corresponding to five imperfect repeats observed in their sequences. Interactions of each of CsgA fragments with all Alpha-synuclein fragments were studied.

Data availability

Datasets used in this study are available at GitHub repository:
https://github.com/KubaWojciechowski/PACT

References

1. Ross, C. A. & Poirier, M. A. Protein aggregation and neurodegenerative disease. *Nat. medicine* **10**, S10–S17 (2004).
2. Haque, E. *et al.* Protein aggregation: A new challenge in type-ii diabetes. *Adv Biotech & Micro* (2017).
3. Yang-Hartwich, Y., Bingham, J., Garofalo, F., Alvero, A. B. & Mor, G. Detection of p53 protein aggregation in cancer cell lines and tumor samples. In *Apoptosis and Cancer*, 75–86 (Springer, 2015).
4. Janson, J. *et al.* Increased risk of type 2 diabetes in alzheimer disease. *Diabetes* **53**, 474–481 (2004).
5. Sims-Robinson, C., Kim, B., Rosko, A. & Feldman, E. L. How does diabetes accelerate alzheimer disease pathology? *Nat. Rev. Neurol.* **6**, 551–559 (2010).
6. Rajput, A., Rozdilsky, B. & Rajput, A. Alzheimer's disease and idiopathic parkinson's disease coexistence. *J. geriatric psychiatry neurology* **6**, 170–176 (1993).
7. Eisenberg, D. S. & Sawaya, M. R. Structural studies of amyloid proteins at the molecular level. *Annu. Rev. Biochem.* **86**, 69–95 (2017).
8. Sipe, J. D. *et al.* Amyloid fibril protein nomenclature: 2010 recommendations from the nomenclature committee of the international society of amyloidosis. *Amyloid* **17**, 101–104 (2010).
9. Lundmark, K., Westermark, G. T., Olsén, A. & Westermark, P. Protein fibrils in nature can enhance amyloid protein a amyloidosis in mice: Cross-seeding as a disease mechanism. *Proc. Natl. Acad. Sci.* **102**, 6098–6102 (2005).
10. Ren, B. *et al.* Fundamentals of cross-seeding of amyloid proteins: an introduction. *J. Mater. Chem. B* **7**, 7267–7282 (2019).
11. Howie, A. J. & Brewer, D. B. Optical properties of amyloid stained by congo red: history and mechanisms. *Micron* **40**, 285–301 (2009).
12. Nielsen, L., Frokjaer, S., Brange, J., Uversky, V. N. & Fink, A. L. Probing the mechanism of insulin fibril formation with insulin mutants. *Biochemistry* **40**, 8397–8409 (2001).
13. Szulc, N. *et al.* Variability of amyloid propensity in imperfect repeats of csga protein of salmonella enterica and escherichia coli. *Int. journal molecular sciences* **22**, 5127 (2021).

14. Shirahama, T. & Cohen, A. S. High-resolution electron microscopic analysis of the amyloid fibril. *The J. cell biology* **33**, 679–708 (1967).

15. Wang, Z. *et al.* Afm and stm study of β-amyloid aggregation on graphite. *Ultramicroscopy* **97**, 73–79 (2003).

16. Gąsior-Głogowska, M. E., Szulc, N. & Szefczyk, M. Challenges in experimental methods. In *Computer Simulations of Aggregation of Proteins and Peptides*, 281–307 (Springer, 2022).

17. Kotulska, M. & Wojciechowski, J. W. Bioinformatics methods in predicting amyloid propensity of peptides and proteins. In *Computer Simulations of Aggregation of Proteins and Peptides*, 1–15 (Springer, 2022).

18. Navarro, S. & Ventura, S. Computational methods to predict protein aggregation. *Curr. Opin. Struct. Biol.* **73**, 102343 (2022).

19. Thompson, M. J. *et al.* The 3d profile method for identifying fibril-forming segments of proteins. *Proc. Natl. Acad. Sci.* **103**, 4074–4078 (2006).

20. Garbuzynskiy, S. O., Lobanov, M. Y. & Galzitskaya, O. V. Foldamyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics* **26**, 326–332 (2010).

21. Gasior, P. & Kotulska, M. Fish amyloid–a new method for finding amyloidogenic segments in proteins based on site specific co-occurence of aminoacids. *BMC bioinformatics* **15**, 1–8 (2014).

22. Walsh, I., Seno, F., Tosatto, S. C. & Trovato, A. Pasta 2.0: an improved server for protein aggregation prediction. *Nucleic acids research* **42**, W301–W307 (2014).

23. Família, C., Dennison, S. R., Quintas, A. & Phoenix, D. A. Prediction of peptide and protein propensity for amyloid formation. *PloS one* **10**, e0134679 (2015).

24. Burdukiewicz, M. *et al.* Amyloidogenic motifs revealed by n-gram analysis. *Sci. reports* **7**, 1–10 (2017).

25. Wojciechowski, J. W. & Kotulska, M. Path-prediction of amyloidogenicity by threading and machine learning. *Sci. Reports* **10**, 1–9 (2020).

26. Louros, N., Orlando, G., De Vleeschouwer, M., Rousseau, F. & Schymkowitz, J. Structure-based machine-guided mapping of amyloid sequence space reveals uncharted sequence clusters with higher solubilities. *Nat. communications* **11**, 1–13 (2020).

27. Kuriata, A. *et al.* Aggrescan3d (a3d) 2.0: prediction and engineering of protein solubility. *Nucleic acids research* **47**, W300–W307 (2019).

28. Szulc, N. *et al.* Bioinformatics methods for identification of amyloidogenic peptides show robustness to misannotated training data. *Sci. reports* **11**, 1–11 (2021).

29. Šali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. molecular biology* **234**, 779–815 (1993).

30. Luca, S., Yau, W.-M., Leapman, R. & Tycko, R. Peptide conformation and supramolecular organization in amylin fibrils: constraints from solid-state nmr. *Biochemistry* **46**, 13505–13522 (2007).

31. Wozniak, P. P. & Kotulska, M. Amyload: website dedicated to amyloidogenic protein fragments. *Bioinformatics* **31**, 3395–3397 (2015).

32. Cani, P. D. Human gut microbiome: hopes, threats and promises. *Gut* **67**, 1716–1725 (2018).

33. Friedland, R. P. & Chapman, M. R. The role of microbial amyloid in neurodegeneration. *PLoS pathogens* **13**, e1006654 (2017).

34. Chen, S. G. *et al.* Exposure to the functional bacterial amyloid protein curli enhances alpha-synuclein aggregation in aged fischer 344 rats and caenorhabditis elegans. *Sci. reports* **6**, 1–10 (2016).

35. Bhoite, S. S., Han, Y., Ruotolo, B. T. & Chapman, M. R. Mechanistic insights into accelerated α-synuclein aggregation mediated by human microbiome-associated functional amyloids. *J. Biol. Chem.* 102088 (2022).

36. Ulamec, S. M. *et al.* Single residue modulators of amyloid formation in the n-terminal p1-region of α-synuclein. *Nat. communications* **13**, 1–16 (2022).

37. Louros, N. *et al.* Mapping the sequence specificity of heterotypic amyloid interactions enables the identification of aggregation modifiers. *Nat. communications* **13**, 1–20 (2022).

38. Burdukiewicz, M. *et al.* Amylograph: a comprehensive database of amyloid–amyloid interactions. *Nucleic Acids Res.* (2022).

39. Nowakowska, A. W. & Kotulska, M. Topological analysis as a tool for detection of abnormalities in protein-protein interaction data. *Bioinformatics* (2022).

40. Ahmed, A. B., Znassi, N., Château, M.-T. & Kajava, A. V. A structure-based approach to predict predisposition to amyloidosis. *Alzheimer's & Dementia* **11**, 681–690 (2015).

41. Hunter, J. D. Matplotlib: A 2d graphics environment. *Comput. science & engineering* **9**, 90–95 (2007).

42. Harris, C. R. *et al.* Array programming with numpy. *Nature* **585**, 357–362 (2020).

43. McKinney, W. *et al.* Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, vol. 445, 51–56 (Austin, TX, 2010).

44. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *J. machine Learn. research* **12**, 2825–2830 (2011).

45. Waskom, M. L. Seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).

Acknowledgements

We would like to thank the team of AmyloGraph developers for their effort in preparing the database used in this project.

We would like to thank Alicja Nowakowska for valuable discussions

This work was partially supported by the National Science Centre, Poland, Grant 2019/35/B/NZ2/03997. Access to Wroclaw Centre for Networking and Supercomputing is greatly acknowledged.

Author contributions statement

J.W.W. and M.K. developed the concept. J.W.W. and W.S. implemented the algorithms. J.W.W. and M.K. analyzed data and wrote the manuscript.