B-SIDER: Computational Algorithm for the Design of Complementary β-sheet Sequences

Tae-Geun Yu[1], *Hak-Sung Kim*[1‡]*, *and Yoonjoo Choi*[1‡]*

[1]Department of Biological Sciences, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea

‡These authors contributed equally.

* Corresponding authors

Email: yoonjoo.choi@kaist.ac.kr

Email: hskim76@kaist.ac.kr

Abstract

The β-sheet is an element of protein secondary structure, and intra-/inter-molecular β-sheet interactions play pivotal roles in biological regulatory processes including scaffolding, transporting, and oligomerization. In nature, a β-sheet formation is tightly regulated because dysregulated β-stacking often leads to severe diseases such as Alzheimer's, Parkinson's, systemic amyloidosis, or diabetes. Thus, the identification of intrinsic β-sheet forming propensities can provide valuable insight into protein designs for the development of novel therapeutics. However, structure-based design methods may not be generally applicable to such amyloidogenic peptides mainly owing to high structural plasticity and complexity. Therefore, an alternative design strategy based on complementary sequence information is of significant importance. Herein, we developed a database search method called B-SIDER for the design of complementary β-strands. This method makes use of the structural database information and generates query-specific score matrices. The discriminatory power of the B-SIDER score function was tested on representative amyloidogenic peptide substructures against a sequence-based score matrix (PASTA2.0) and two popular *ab initio* protein design score functions (Rosetta and FoldX). B-SIDER is able to distinguish wild-type amyloidogenic β-strands as favored interactions in a more consistent manner than other methods. B-SIDER was prospectively applied to the design of complementary β-strands for a splitGFP scaffold. Three variants were identified to have stronger interactions than the original sequence selected through a directed evolution, emitting higher fluorescence intensities. Our results indicate that B-SIDER can be applicable to the design of other β-strands, assisting in the development of therapeutics against disease-related amyloidogenic peptides.

Introduction

The β-sheet is one of the major units of protein structure [1], and has a variety of functions in transportation, recognition, scaffolding, and enzymatic processes [2]. The mechanism of a β-sheet formation has recently received significant attention owing to its close relations with several critical diseases such as Alzheimer's, Parkinson's, type 2 diabetes, and systemic amyloidosis [3, 4]. Such diseases are known to be linked to the precipitation of dysregulated β-stacking between neighboring β-strands [5-7]. In this regard, information on the amino acid propensity of intrinsic β-sheet forming motifs and its use in the design of their complementary sequences are crucial for understanding the mechanism of β-sheet formation and developing potential therapeutics specifically targeting aggregation-prone regions [8].

Whereas structure-based protein design approaches have shown notable successes in several cases [9], their application to *de novo* β-sheet designs still remains a challenge [2, 10]. Although structure-based design approaches require a well-defined protein structure, amyloidogenic peptides usually have highly disordered structures [11]. The structural identification of such peptides has long been hindered by high degrees of structural plasticity, transiency, and complexity owing to a self-oligomerization [12, 13]. It is thus necessary to exploit the complementarity across neighboring β-strand pairs using sequence information.

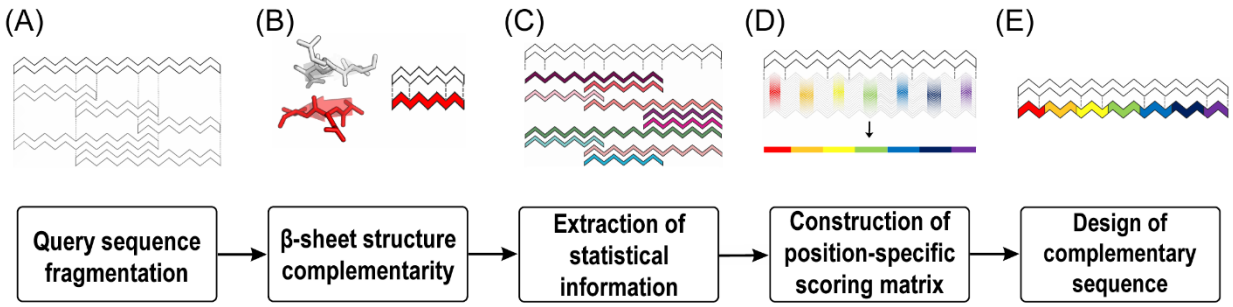
Intriguingly, significant conservation and covariations of residue pairs between neighboring β-strands have been identified in many different protein families [14]. For instance, pairs of β-branched residues and cysteines are preferred at nonhydrogen-bonded positions. Aromatic residues tend to be paired with valine or glycine [15]. Several computational algorithms have been developed to predict aggregation-prone regions based on the internal β-sheet forming patterns. While differing in detail, they make use of either statistical potentials such as Tango [16], PASTA [17], SALSA [18],

BETASCAN [19], and Waltz [20] or physicochemical properties of amino acids [21]. In addition, consensus methods and machine-learning approaches have also been developed [22, 23], showing fine agreement with the experiment results.

It has been reported that β-strand interactions can be stabilized by introducing β-sheet favored pairs [24-28] and charge pairing between neighboring β-strands [29-31]. Recent studies have shown that fragments derived from the amyloidogenic region can be used for β-stack modeling [6, 32]. While the use of amino acid pairing information in a protein design has been attempted elsewhere, practical applications of such patterns have been limited, mainly owing to the lack of a comprehensive quantification for residue pairing and noisy patterns of β-sheet forming residue pairs [1, 33, 34]. The β-sheet forming peptides appear to have poor sequence commonalities and imperfect repeats [19]. Therefore, a careful curation of meaningful patterns is required for a practical protein design strategy of complementary β-strands.

Herein, we present a database search method, B-SIDER (β-Stacking Interaction DEsign for Reciprocity), for the design of complementary β-strands. The method generates a query-specific score matrix from the structure database. To utilize the pairing information and overcome the pattern noise, we hypothesized that significant complementary pairs can be amplified by superposing a subset of sequence fragments. Moreover, the recent growth boom of β-sheet structures [35] allows the solid statistics of β-sheet forming residue pairings [36]. Based on the hypothesis and statistics of β-sheet forming residue pairings, we developed a fast and reliable computational method for the design of complementary β-strand sequences. The methodology augments β-sheet forming residue preferences by overlaying complementary fragment sequences (**Figure 1**). We retrospectively validated our approach using a set of curated amyloidogenic targets and compared it with two popularly used structure-based methods (Rosetta and FoldX) and a

82 sequence-based aggregation prediction algorithm (PASTA 2.0). Our algorithm was shown to
83 clearly distinguish favorable β-sheet forming sequences entirely based on the query sequence,
84 whereas the structure-based energy functions exhibited inconsistent results depending on the
85 targets. The utility and potential of our method were demonstrated by designing novel
86 complementary peptides for splitGFP. The designed sequences showed stronger interactions with
87 neighboring strands of the scaffold, and consequently higher fluorescence emissions, than the
88 original peptide selected through directed mutagenesis [37].



89 **Figure 1. Overview of the B-SIDER algorithm.** (A) When a query sequence is given, the query
90 is divided into smaller linear peptides ranging from three residues to its full length. (B) Exactly
91 matched sequences are identified and checked against the structure database such that the matches
92 indeed form β-sheets. (C) If the matches form β-sheets, their complementary sequences are
93 extracted. (D) A position-specific score matrix is constructed based on the complementary
94 sequence information. (E) Final complementary sequences are designed using the score matrix.

95
96
97

Materials and Methods

Computational algorithm for the design of complementary sequences

Collection of β-strand information Non-redundant structures determined by high-resolution X-ray crystallography were collected from the PISCES [38] (10~90% sequence identity), < 3 Å resolution, < 0.3 R-factor and sequence length from 40 to 10000. Given the query target sequence, the non-redundant structure database was used to extract pairing information from matched sequences with the same directionality. The target sequence is initially divided into linear moving-windows whose residues in length range from 3 to the entire target-sequence length. Any structures with identical target sequence fragments as the split queries were collected, followed by further filtering based on the definition of β-sheet secondary structure (the distance between the backbone nitrogen-oxygen atom pairs < 5 Å). To remove any redundancy, protein structures that contain matches were compared using TMalign [39]. If the TM-score > 0.7 and sequence identity > 90%, one of the matched sequences is removed.

While the method is applicable to both parallel and anti-parallel β-sheets in theory, we entirely focused on anti-parallel β-sheets in this study [27, 28] and all the test cases are anti-parallel because anti-parallel cases are more frequently observed compared to parallel cases [40]. Disease-related amyloidogenesis is also known to be initiated with anti-parallel β-sheets, and soluble oligomeric amyloid species mainly exist as anti-parallel [41, 42].

Complementary sequence score The β-sheet complementarity score function is derived from the environment-specific substitution score [43]. We hypothesized that each position of a β-strand is independent of one another, and their complementarities are determined through residue pairs from

neighboring strands. Given the query sequence, all identified neighboring sequences are pooled together, as described in the previous section. The amino acid frequency at each complementary position is counted as

$$A_{i,p} = O_{i,p} / \sum_i O_{i,p}. \quad (1)$$

where $A_{i,p}$ is the frequency of a certain amino acid (i) at a specific complementary position (p), and $O_{i,p}$ is the total count of the amino acid at p. The background frequency of a certain amino acid B_i is independent of the query and thus counted from a well defined structure database(HOMSTRAD database [44]). The frequency is calculated as

$$B_i = \sum_p O_{i,p} / \sum_{i,p} O_{i,p}. \quad (2)$$

The complementary sequence score of the amino acid at the position $S_{i,p}$ is calculated as

$$S_{i,p} = -\log(A_{i,p}/B_i). \quad (3)$$

It should be noted that the complementarity score is completely data-driven, i.e., if an amino acid never appears at a certain position, a high penalty score is imposed. We only consider complementary amino acids that are found at least once in the entire identified sequences. The final score is the sum of the scores at all positions.

Protein expression and complementation assay

Gene construction The gene coding for splitGFP [37] consists of the first through tenth (GFP1-10) and 11th strand (GFP11) templates (Supporting Information, **Table S1**). They were cloned into pET-28a (Novagen) vector between the Nde-I and Xho-I restriction sites. We introduced additional mutations to GFP1-10 to inhibit aggregation and for a convenient expression [45]. The GFP11 strand

was fused with a P22 virus-like particle scaffolding protein [46] for a soluble and stable expression. Mutations on GFP11 were introduced through PCR using mutagenic primers (Supporting Information, **Table S2**), and the resulting genes were cloned into the pET-28a vector. Six histidine residues were fused to the N-terminal of the GFP1-10 and GFP11 genes as an affinity purification tag.

Protein expression and purification All constructs were transformed into BL21 (DE3) Escherichia coli strains. The transformed cells were grown overnight and inoculated into a Luria-Bertani media containing 50 μg/ml of kanamycin at 37 ℃. The cells were then grown until the optical density of the cells reached 0.6–0.8 at 600 nm, followed by the addition of 0.7 mM of IPTG (isopropyl β-D-1-thiogalactopyranoside) for induction. After incubation for 16–18 h at 18 ℃, the cells were harvested and suspended in a lysis buffer containing 50 mM Tris (pH 8.0), 150 mM NaCl, and 5 mM imidazole. The suspended cells were disrupted through sonication, and insoluble fractions were removed using centrifugation at 18,000 g for 1 h. The supernatants were filtered with 0.22 μm syringe filters and purified through affinity chromatography using Ni-NTA agarose Superflow (Qiagen). The solutions were applied to the resin-packed columns and washed with a buffer containing 50 mM Tris (pH 8.0), 150 mM NaCl, and 10 mM imidazole, until no proteins were detected through a Bradford assay. An elution buffer (50 mM Tris (pH 7.4), 150 mM NaCl, 300 mM imidazole) was then applied to the column. The buffer exchange was performed with phosphate buffered saline (PBS, pH 7.4) using a PD-10 column (GE health-care). The concentrations of the proteins were determined by measuring the absorbance at 280 nm. All purification processes were conducted at 4 ℃. The purities of the proteins were then evaluated using SDS-PAGE.

Complementation assay The assembly of splitGFP variants was monitored and measured using a fluorescence complementation assay. An excessive amount of GFP1-10 (50 pmol) in 180 µl and 20 µl of equal molar concentration for each GFP 11 strand (3 pmol) were co-incubated in a PBS buffer (pH 7.4). The fluorescence kinetics (λ_{ex}= 488 nm /λ_{em}= 530 nm) were monitored for 12 h at 25 ℃ using a TECAN infinite M200 microplate reader at 5 min intervals [37] with shaking for 2 sec between intervals. Each experiment was conducted in triplicate using a Nunc F 96 Micro-well black plate, blocked with a solution of PBS containing 0.5 % of bovine serum albumin (BSA) for 30 min before the assay.

Results/Discussion

Overview of the design process

We hypothesized that repetitively observed amino acid pairing patterns indicate a strong preference toward the β-sheet. It was also assumed that the sequence with the most frequent patterns will directly form a β-sheet without considering other environmental contributions.

There are two major steps applied in the algorithm: 1) The extraction of complementarity information of the β-sheet and 2) the construction of a scoring matrix (**Figure 1**). When a query sequence is given, it is fragmented into several pieces of short peptides longer than three residues in length, and the matching neighboring strands are collected. These fragmentation and overlaying processes naturally impose weights on complementary-prone positions and amplify the pattern signals. It should be noted that the subsequence search starts from the -2nd position in order to avoid underweighting teminal regions (Supporting Information, **Figure S1**). After the collection

183 of matched sequences, a position-specific complementarity scoring matrix is constructed. The
184 scoring matrix obtained is used to evaluate and design the complementarity of β-strand interactions.

185 **Validation of the score function on retrospective cases**

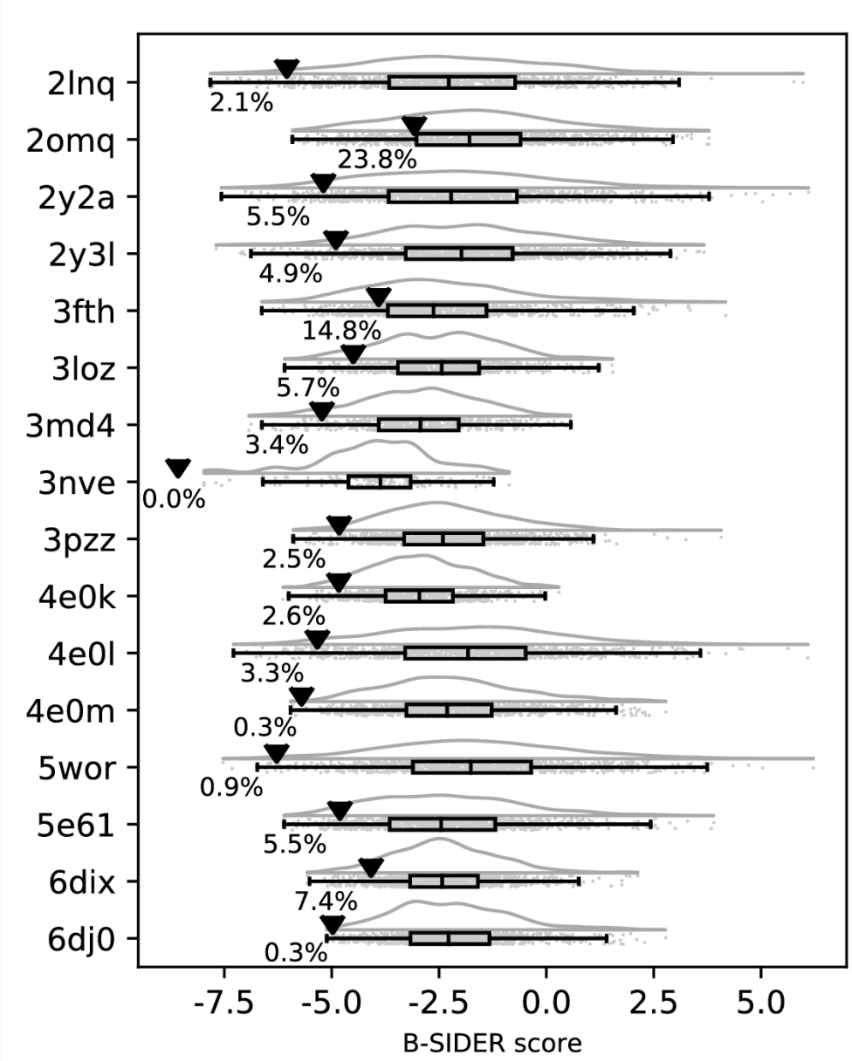
186 In an effort to validate the complementarity score, we manually curated a test set of
187 naturally occurring β-strand pairs whose environmental effects are minimal. It is known that β-
188 strand pairing is in general significantly hindered by local environments [47], whereas amyloidogenic
189 peptide segments are known to form natural β-sheets themselves [17]. We thus selected a set of
190 widely known amyloidogenic structures whose aggregation-prone regions have been identified
191 (**Table1** and Supporting Information, **Table S3**). When multiple β-strands are present, we assumed
192 that the first strands may be the primary amyloid building unit and so only the first two strands as
193 a pair were extracted for the structure-based energy calculation.

194 **Table 1. Retrospective test set**

Source	PDB ID
Amyloid-β	2lnq, 2y2a, 2y3l, 3pzz
Insulin	2omq
IAPP(amylin)	3fth, 5e61
Prion	3md4, 3nve
Tau	4e0m
β-2 microglobulin	3loz, 4e0k, 4e0l
Immunoglobulin	6dj0, 6dix
SOD1	5wor

195

196 To assess the complementarity of the native sequences, we compared their scores with
197 those of random sequences which were generated using all 20 amino acids in a position-
198 independent manner. Natural amyloidogenic segments are known to be highly prone to



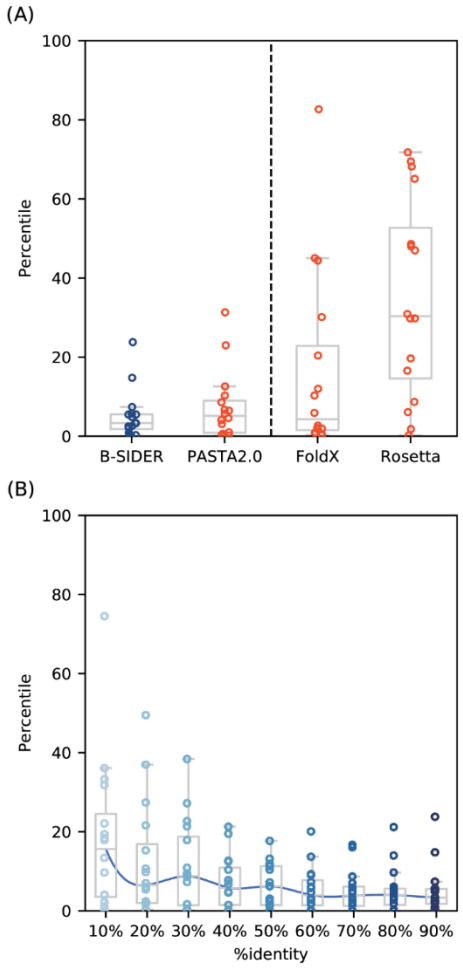
199 aggregation, and thus they are expected to be highly preferred, i.e., having fairly low scores in the
200 random sequence score distributions. **Figure 2** shows that all native sequence scores are ranked
201 extremely low in all distributions. On average, the native sequences are within 5.2% of the
202 distributions (**Figure 2**). The results indicate that the scoring function is extremely useful in
203 detecting favorable β-strand counterparts.

Figure. 2. The predictive power of the B-SIDER score. In this test set, native sequences were compared against 1,000 random sequences. A lower score indicates more favorability. The native complementarity scores are marked as (▼) and their percentile values are displayed next to the marks. This plot was generated using the Raincloud Python package [48].

We also compared the B-SIDER score with two structure-based all-atom energy functions and a sequence-based score matrix. For the structure-based methods, we picked Rosetta (Talaris 13) [49, 50] and FoldX [51], which have been popularly used in *de novo* protein designs [52, 53]. PASTA 2.0 [54] is a method for predicting regions prone to aggregation using a scoring matrix derived from residue pairing patterns of β-sheets. To avoid any biases, 1,000 random sequences were newly prepared per target. The "FastRelax" protocol [55] from Rosetta (Ver. 3.7), the "BuildModel" command from FoldX (Ver. 4.0), and the scoring matrix from PASTA 2.0 were used against the native and random sequences. The predictive power of the score function was assessed based on the percentile of the native sequence score against the random sequence score distribution.

Figure 3A shows that structure-based score functions are generally worse than the sequence-based scoring matrices. The results indicate that the Rosetta energy score function is not sufficiently accurate for ranking complementary β-strands (35.1th percentile on average), whereas the predictive powers of PASTA 2.0 and FoldX were moderate, showing 7.4th and 16.4th percentiles on average respectively. B-SIDER was shown to be the most accurate in an extremely consistent manner (5.2th on average. Standard deviations of Rosetta, FoldX, PASTA 2.0 and B-SIDER are 25.2, 23.3, 8.7 and 6.2, respectively). Although the assessment by PASTA 2.0 is also fairly consistent, the query-specific nature of B-SIDER may provide better results.

226



227 **Figure. 3. Predictive power of B-SIDER score and its robustness.** (A) The B-SIDER score is
228 compared against two popularly used structure-based protein design methods (Rosetta and FoldX)
229 and the scoring matrix for the prediction of an amyloid formation (PASTA 2.0). For a fair
230 comparison, we generated new sets of 1,000 random sequences per target and assessed their scores
231 using each energy function. The percentile values of native complementary sequences against the
232 score distributions of the randomly generated sequences are presented. The B-SIDER score
233 provided the most consistent assessment. (B) The robustness of the B-SIDER score was evaluated
234 at various sequence identity threshold values. The percentile values of B-SIDER are overall fairly
235 consistent at even low sequence identity cutoff.

Considering that the Rosetta relax protocol performs flexible backbone refinements, the use of a fixed-backbone calculation seems to be better for an evaluation of the β-sheet complementarity. It should be noted that the inconsistent results of Rosetta imply that FoldX prediction will also be highly driven by the preparation of the structure, i.e., a design with an ill-defined model may not be generally successful. By contrast, B-SIDER and PASTA 2.0 do not require any structure at all, and thus can be applied to general cases such as β-sheet interactions with high structural plasticity and poor structural integrity, which are the common features of amyloidogenic peptides. Furthermore, the process of collecting complementary motifs of B-SIDER also appears to be powerful, making it possible to distinguish favorable complementary sequences not easily detected through one-to-one residue pairing. We also tested the robustness of the B-SIDER approach by examining the algorithm at various sequence identity cutoff values of the structure database. The results are fairly consistent until < 40% sequence identity cutoff (**Figure 3B** and Supporting Information, **Figure S2)**. We observe that significant outliers start appearing at < 30%. Despite such outliers, B-SIDER still successfully discriminates most test cases as favorable. The length distributions of matched subsequences (**Figure S3**) show that complementarity information is mainly derived from subsequences with the minimal length. Perhaps the consistency of the predictive power at low homology thresholds may come from the subsequence overlapping.

Prospective appplication of the algorithm to splitGFP

As shown in the retrospective test, B-SIDER is extremely useful in discriminating naturally β-strand forming sequences. As a proof of concept, we prospectively designed novel complementary β-strands for splitGFP. SplitGFP is a fragmented protein pair derived from

superfolderGFP [37], comprising a scaffold containing ten β-strands (GFP1-10) and their complementary β-strand peptide (GFP11). GFP11 specifically interacts with GFP1-10, and the strand tightly forms a stable β-sheet structure, which facilitates the chromophore maturation in an irreversible manner [56]. This assembly process results in the emission of green fluorescence. Because GFP11 is known to be disordered in a solution, its conformational transition from a disordered to an induced β-sheet is similar to that of amyloidogenic peptides [57, 58]. This model system thus efficiently assesses whether sequences designed using B-SIDER have favorable β-sheet interactions.

Table 2. GFP 11 variants

Variant	Sequence	Score	Relative Assembly*
top_var1	M**V**L**V**E**F**V**T**	-12.34	1.17
top_var2	M**Y**L**V**E**F**V**T**	-12.14	1.4
top_var9	M**T**L**V**E**F**V**T**	-11.79	1.36
top_var3	M**V**L**V**E**I**V**T**	-12.11	0.59
top_var4	M**V**L**V**E**F**V**Y**	-11.95	0.56
top_var5	M**Y**L**V**E**I**V**T**	-11.92	0.88
top_var6	M**V**L**V**E**V**V**T**	-11.90	N.D.
top_var7	M**V**L**V**E**F**V**V**	-11.89	N.D.
top_var8	M**V**L**V**E**F**V**W**	-11.83	N.D.
top_var10	M**V**L**V**E**F**V**F**	-11.77	N.D.
neg_var	M**V**L**G**E**K**V**E**	-4.60	0.04
Ori	MVLHEYVN	-6.50	1

*Normalized values compared to the fluorescence level of the original GFP11 strand.

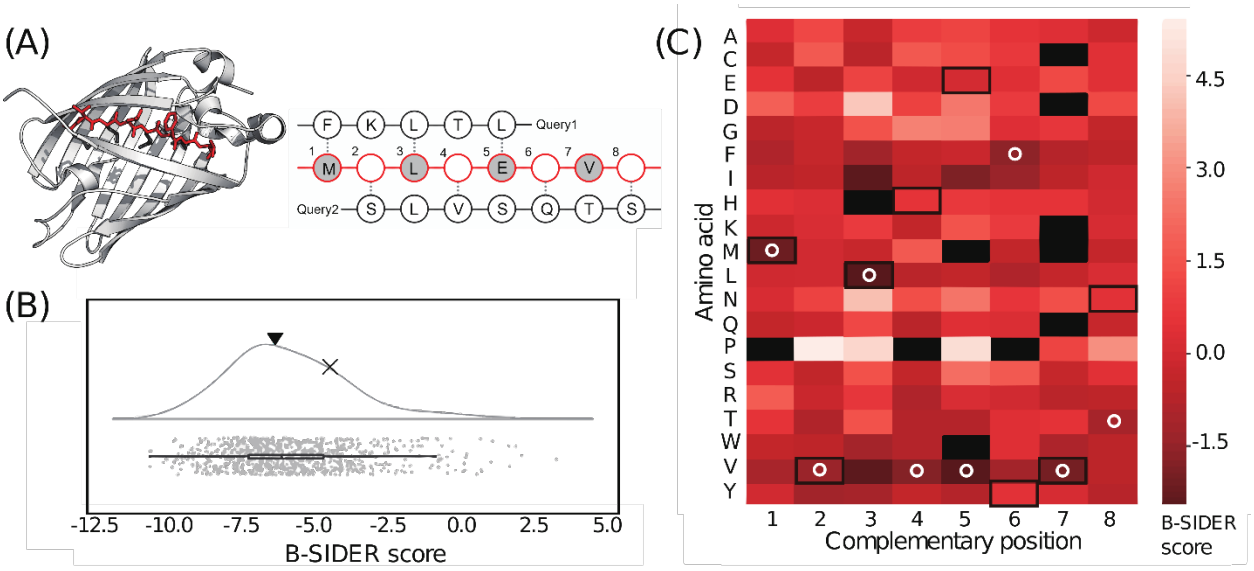


Figure. 4. Design process of GFP11 variants. (A) The strand to be designed (GFP11) is highlighted in red, and query sequences to consider are shown on the right. Dotted lines represent hydrogen bonds. The residues pointing inward, which are not subject to a mutation, are colored in gray. (B) The B-SIDER score of the original GFP11 (▼) is compared against the score distribution of randomly generated sequences. The negative variant (neg_var) is marked as × (the 77th percentile). (C) The position-specific scoring matrix for the query sequences. Amino acids that never appeared at each position are colored in black. The amino acids of the native sequence are highlighted with black boxes. The white circles represent the lowest scores at each position.

The original GFP11 was designed using directed mutagenesis, and it shows a high intrinsic propensity to form hydrogen bonds with the neighboring β-strands of GFP1-10 [59]. In our case, the queries are the neighboring strands of GFP1-10 (**Figure 4A**). The total score was calculated as a sum of the scores from both sides without specific weights. It is known that the inward pointing residues (1, 3, 5, and 7th positions) directly interact with the chromophore, and thus are not subject to a mutation. It should be noted that we utilized the best structure data available (PDB < 90 %

276 sequence identity). B-SIDER identified 2,637 non-redundant sequences from the structure
277 database. The native sequence is ranked modestly among 1,000 possible randomly chosen
278 sequence variants (46th percentile), indicating that there could be room for complementary
279 sequences with stronger interactions than the original (**Figure 4B**). We then selected sequences
280 with the lowest B-SIDER scores (top_vars; **Table2**). Amino acid compositions of the ten variants
281 are mostly hydrophobic or branched amino acids (Supporting Information, **Figure S4**). In addition,
282 one sequence with a high score (> 75th percentile) was randomly selected as a negative control
283 (neg_var, 77th).

284 The selected variants were successfully expressed and purified (Supporting Information,
285 **Figure S5**) except for four clones (top_var6, 7, 8, and 10), which were observed to be insoluble,
286 perhaps due to aggregation. Among those expressed, three variants (top_var1, 2, and 9) showed
287 faster assembly patterns and higher signals compared to the original GFP11 (**Figure 5**). No
288 functional aberrance with excitations or emissions was observed (Supporting Information, **Figure**
289 **S6**). All successful variants, which emitted stronger fluorescence levels than the original, were
290 shown to have pairs of phenylalanine and threonine at positions 6 and 8, respectively. These results
291 demonstrate that the designed variants indeed formed complementary β-strands in a more
292 favorable fashion than the original peptide as predicted. The other variants showed slightly lower
293 signals than the original, but still gave rise to clear fluorescence signals (**Figure 5**). The negative
294 control (neg_var) barely emitted any signal, suggesting that the score indeed indicates the
295 complementarity of the β-stacking interactions. We also assigned scores to the GFP11 variants
296 using other scoring methods. As shown in the retrospective test set, Rosetta and FoldX were unable
297 to discriminate top_vars as favorable (Supporting Information, **Figure S7**). However, PASTA 2.0
298 was again fairly accurate in this case.

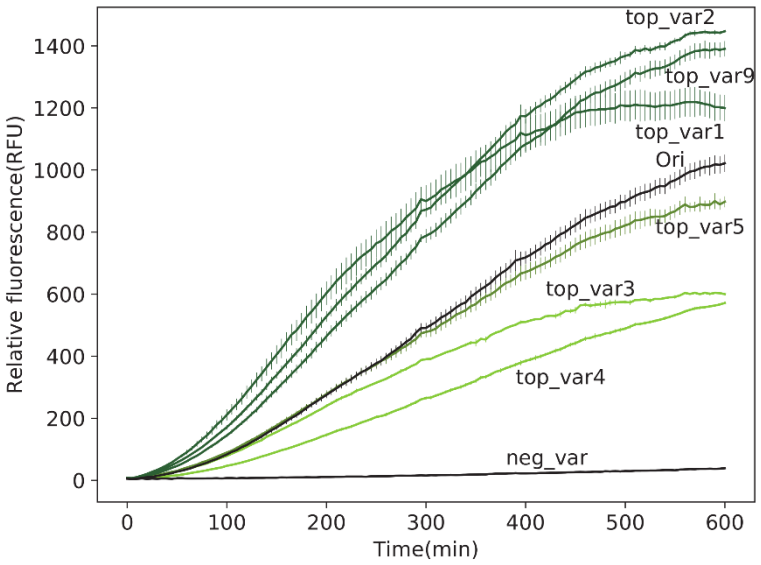


Figure. 5. Complementation assay with the designed GFP11 variants. Among the six successfully expressed variants, three variants exhibited stronger fluorescence emissions than the original peptide identified through a directed mutagenesis.

Conclusion

For understanding the aggregation mechanism of disease-related β-sheets and developing potential therapeutics against them, β-sheet forming patterns are essential. Unlike α-helices, however, there has been no established design principle for the complementarity of β-sheets. In this study, we developed B-SIDER, a database search method for the design of complementary β-strands based on the intrinsic β-sheet forming propensities. Statistical patterns of interacting residue pairs between neighboring β-strands enable the complementary interaction to be quantified. We demonstrated that the statistical potential can be directly applied to the design of complementary β-strand sequences. Using splitGFP as a model system, we successfully designed fragment variants, which led to stronger fluorescence emissions than the native one originally

identified through a directed mutagenesis. The results clearly indicate that B-SIDER can be useful for the detection and design of β-stacking interactions between unstructured fragments. Therefore, our approach can find wide applications in protein designs where structure-based methods are ineffective, including the development of protein binders specifically against intrinsically disordered disease-related proteins.

Notes

B-SIDER is freely available at http://bel.kaist.ac.kr/research/B-SIDER

Supporting Information

Supporting Information Available: Sequences of GFP constructs (Table S1 and S2), identification of query sequences of the retrospective test set (Table S3), schematic description of B-SIDER scoring (Figure S1), discriminatory power of B-SIDER scoring at various sequence identity cutoff values (Figure S2), length distribution of matched subsequences (Figure S3), sequence logo of GFP11 variants (Figure S4), SDS-PAGE analysis of GFP11 variants (Figure S5), excitation/emission spectral analyses of GFP11 variants (Figure S6), score reassessment of the GFP11 variants using PASTA2.0, FoldX and Rosetta (Figure S7).

This material is available free of charge via the Internet at http://pubs.acs.org

Acknowledgements

This work was conducted using Alphacom high-performance computing cluster at the department of biological sciences at the Korea Advanced Institute of Science and Technology (KAIST).

This work was supported by the Korea Research Fellowship Program (2016H1D3A1938246 to Y.C.), Global Research Laboratory (NRF-2015K1A1A2033346), and Mid-Career Researcher Program (NRF-2017R1A2A1A05001091) of the National Research Foundation (NRF) funded by the Ministry of Science and ICT.

Abbreviations

GFP, green fluorescent protein; PCR, polymerase chain reaction; IPTG, isopropyl β-D-1-thiogalactopyranoside; BSA, bovine serum albumin; IAPP, islet amyloid polypeptide precursor; SOD1, superoxide dismutase 1; PDB, protein data bank

References

1. Bhattacharjee, N.; Biswas, P., Position-Specific Propensities of Amino Acids in the β-strand. *BMC Struct. Biol.* **2010**, 10, 29.

2. Marcos, E.; Chidyausiku, T. M.; McShan, A. C.; Evangelidis, T.; Nerli, S.; Carter, L.; Nivón, L. G.; Davis, A.; Oberdorfer, G.; Tripsianes, K.; others, De novo Design of a Non-local β-sheet Protein with High Stability and Accuracy. *Nat. Struct. Mol. Biol.* **2018**, 25, 1028.

3. Chiti, F.; Dobson, C. M., Protein Misfolding, Amyloid Formation, and Human Disease: a Summary of Progress Over the Last Decade. *Annu. Rev. Biochem.* **2017**, 86, 27-68.

4. Richardson, J. S.; Richardson, D. C., Natural β-sheet Proteins Use Negative Design to Avoid Edge-to-edge Aggregation. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, 99, 2754-2759.

5. Colletier, J.-P.; Laganowsky, A.; Landau, M.; Zhao, M.; Soriaga, A. B.; Goldschmidt, L.; Flot, D.; Cascio, D.; Sawaya, M. R.; Eisenberg, D., Molecular Basis for Amyloid-β Polymorphism. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, 108, 16938-16943.

6. Liu, C.; Zhao, M.; Jiang, L.; Cheng, P.-N.; Park, J.; Sawaya, M. R.; Pensalfini, A.; Gou, D.; Berk, A. J.; Glabe, C. G.; others, Out-of-register β-sheets Suggest a Pathway to Toxic Amyloid Aggregates. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, 109, 20913-20918.

7. Matthes, D.; Daebel, V.; Meyenberg, K.; Riedel, D.; Heim, G.; Diederichsen, U.; Lange, A.; de Groot, B. L., Spontaneous Aggregation of the Insulin-derived Steric Zipper Peptide VEALYL Results in Different Aggregation Forms with Common Features. *J. Mol. Biol.* **2014**, 426, 362-376.

8. Giorgetti, S.; Greco, C.; Tortora, P.; Aprile, F., Targeting Amyloid Aggregation: an Overview of Strategies and Mechanisms. *Int. J. Mol. Sci.* **2018**, 19, 2677.

9. Huang, P.-S.; Boyken, S. E.; Baker, D., The Coming of Age of De novo Protein Design. *Nature* **2016**, 537, 320.

10. Dou, J.; Vorobieva, A. A.; Sheffler, W.; Doyle, L. A.; Park, H.; Bick, M. J.; Mao, B.; Foight, G. W.; Lee, M. Y.; Gagnon, L. A.; others, De novo Design of a Fluorescence-activating β-barrel. *Nature* **2018**, 561, 485.

11. Jang, H.; Arce, F. T.; Lee, J.; Gillman, A. L.; Ramachandran, S.; Kagan, B. L.; Lal, R.; Nussinov, R. Computational Methods for Structural and Functional Studies of Alzheimer's Amyloid Ion Channels. In *Protein Amyloid Aggregation*; Springer: 2016, pp 251-268.

12. Dovidchenko, N. V.; Galzitskaya, O. V. Computational Approaches to Identification of Aggregation Sites and the Mechanism of Amyloid Growth. In *Lipids in Protein Misfolding*; Springer: 2015, pp 213-239.

13. Zheng, W.; Tsai, M.-Y.; Chen, M.; Wolynes, P. G., Exploring the Aggregation Free Energy Landscape of the Amyloid-β Protein (1-40). *Proc. Natl. Acad. Sci. U.S.A.* **2016**, 113, 11835-11840.

14. Mandel-Gutfreund, Y.; Zaremba, S. M.; Gregoret, L. M., Contributions of Residue Pairing to β-sheet Formation: Conservation and Covariation of Amino Acid Residue Pairs on Antiparallel β-strands. *J. Mol. Biol.* **2001**, 305, 1145-1159.

15. Steward, R. E.; Thornton, J. M., Prediction of Strand Pairing in Antiparallel and Parallel β-sheets Using Information Theory. *Proteins* **2002**, 48, 178-191.

16. Fernandez-Escamilla, A.-M.; Rousseau, F.; Schymkowitz, J.; Serrano, L., Prediction of Sequence-dependent and Mutational Effects on the Aggregation of Peptides and Proteins. *Nat. Biotechnol.* **2004**, 22, 1302.

17. Trovato, A.; Chiti, F.; Maritan, A.; Seno, F., Insight into the Structure of Amyloid Fibrils from the Analysis of Globular Proteins. *PLOS Comput. Biol.* **2006**, 2, 170.

18. Zibaee, S.; Makin, O. S.; Goedert, M.; Serpell, L. C., A Simple Algorithm Locates β-strands in the Amyloid Fibril Core of α-synuclein, Aβ, and Tau Using the Amino Acid Sequence Alone. *Protein Sci.* **2007**, 16, 906-918.

19. Bryan Jr, A. W.; Menke, M.; Cowen, L. J.; Lindquist, S. L.; Berger, B., BETASCAN: Probable β-Amyloids Identified by Pairwise Probabilistic Analysis. *PLOS Comput. Biol.* **2009**, 5, 1000333.

20. Maurer-Stroh, S.; Debulpaep, M.; Kuemmerer, N.; De La Paz, M. L.; Martins, I. C.; Reumers, J.; Morris, K. L.; Copland, A.; Serpell, L.; Serrano, L.; others, Exploring the Sequence Determinants of Amyloid Structure Using Position-specific Scoring Matrices. *Nat. Methods* **2010**, 7, 237.

21. Tartaglia, G. G.; Vendruscolo, M., The Zyggregator Method for Predicting Protein Aggregation Propensities. *Chem. Soc. Rev.* **2008**, 37, 1395-1401.

22. Kim, C.; Choi, J.; Lee, S. J.; Welsh, W. J.; Yoon, S., NetCSSP: Web Application for Predicting Chameleon Sequences and Amyloid Fibril Formation. *Nucleic Acids Res.* **2009**, 37, 469.

23. Tsolis, A. C.; Papandreou, N. C.; Iconomidou, V. A.; Hamodrakas, S. J., A Consensus Method for the Prediction of 'Aggregation-prone' Peptides in Globular Proteins. *PLOS One* **2013**, 8, 54175.

24. Kortemme, T.; Ramı´rez-Alvarado, M.; Serrano, L., Design of a 20-amino Acid, Three-Stranded β-sheet Protein. *Science* **1998**, 281, 253-256.

25. Minor Jr, D. L.; Kim, P. S., Measurement of the β-sheet-forming Propensities of Amino Acids. *Nature* **1994**, 367, 660.

26. Quinn, T. P.; Tweedy, N. B.; Williams, R. W.; Richardson, J. S.; Richardson, D. C., Betadoublet: De novo Design, Synthesis, and Characterization of a Beta-sandwich Protein. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, 91, 8747-8751.

27. Stranges, P. B.; Machius, M.; Miley, M. J.; Tripathy, A.; Kuhlman, B., Computational Design of a Symmetric Homodimer Using β-strand Assembly. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, 108, 20562-20567.

28. Hu, X.; Wang, H.; Ke, H.; Kuhlman, B., Computer-based Redesign of a β sandwich Protein Suggests That Extensive Negative Design Is Not Required for De novo β sheet Design. *Structure* **2008**, 16, 1799-1805.

29. Shammas, S. L.; Knowles, T. P.; Baldwin, A. J.; MacPhee, C. E.; Welland, M. E.; Dobson, C. M.; Devlin, G. L., Perturbation of the Stability of Amyloid Fibrils Through Alteration of Electrostatic Interactions. *Biophys. J.* **2011**, 100, 2783-2791.

30. Wang, W.; Hecht, M. H., Rationally Designed Mutations Convert De novo Amyloid-like Fibrils into Monomeric β-sheet Proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, 99, 2760-2765.

31. West, M. W.; Wang, W.; Patterson, J.; Mancias, J. D.; Beasley, J. R.; Hecht, M. H., De novo Amyloid Proteins from Designed Combinatorial Libraries. *Proc. Natl. Acad. Sci.* **1999**, 96, 11211-11216.

32. Gallardo, R.; Ramakers, M.; De Smet, F.; Claes, F.; Khodaparast, L.; Khodaparast, L.; Couceiro, J. R.; Langenberg, T.; Siemons, M.; Nyström, S.; others, De novo Design of a Biologically Active Amyloid. *Science* **2016**, 354, 4949.

33. Fujiwara, K.; Toda, H.; Ikeguchi, M., Dependence of α-helical and β-sheet Amino Acid Propensities on the Overall Protein Fold Type. *BMC Struct. Biol.* **2012**, 12, 18.

34. Hutchinson, E. G.; Sessions, R. B.; Thornton, J. M.; Woolfson, D. N., Determinants of Strand Register in Antiparallel β-sheets of Proteins. *Protein Sci.* **1998**, 7, 2287-2300.

35. Marcos, E.; Silva, D.-A., Essentials of De novo Protein Design: Methods and Applications. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2018**, 8, 1374.

36. Sormanni, P.; Aprile, F. A.; Vendruscolo, M., Rational Design of Antibodies Targeting Specific Epitopes Within Intrinsically Disordered Proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, 112, 9902-9907.

37. Cabantous, S.; Terwilliger, T. C.; Waldo, G. S., Protein Tagging and Detection with Engineered Self-assembling Fragments of Green Fluorescent Protein. *Nat. Biotechnol.* **2005**, 23, 102.

38. Wang, G.; Dunbrack Jr, R. L., PISCES: a Protein Sequence Culling Server. *Bioinformatics* **2003**, 19, 1589-1591.

39. Zhang, Y.; Skolnick, J., TM-align: a Protein Structure Alignment Algorithm Based on the TM-score. *Nucleic Acids Res.* **2005**, 33, 2302-2309.

40. Hubbard, T. J. Use of β-strand Interaction Pseudo-Potentials in Protein Structure Prediction and Modeling. In Proceedings of the Hawaii International Conference on System Sciences, 1994; 1994; Vol. 27; pp 336-336.

41. Cerf, E.; Sarroukh, R.; Tamamizu-Kato, S.; Breydo, L.; Derclaye, S.; Dufrêne, Y. F.; Narayanaswami, V.; Goormaghtigh, E.; Ruysschaert, J.-M.; Raussens, V., Antiparallel β-Sheet: a Signature Structure of the Oligomeric Amyloid-β Peptide. *Biochem. J.* **2009**, 421, 415-423.

42. Gordon, D. J.; Balbach, J. J.; Tycko, R.; Meredith, S. C., Increasing the Amphiphilicity of an Amyloidogenic Peptide Changes the β-sheet Structure in the Fibrils From Antiparallel to Parallel. *Biophys. J.* **2004**, 86, 428-434.

43. Choi, Y.; Deane, C. M., FREAD Revisited: Accurate Loop Structure Prediction Using a Database Search Algorithm. *Proteins* **2010**, 78, 1431-1440.

44. Mizuguchi, K.; Deane, C. M.; Blundell, T. L.; Overington, J. P., HOMSTRAD: a Database of Protein Structure Alignments for Homologous Families. *Protein Sci.* **1998**, 7, 2469-2471.

463 45. Kim, Y. E.; Kim, Y.-n.; Kim, J. A.; Kim, H. M.; Jung, Y., Green Fluorescent Protein
464 Nanopolygons as Monodisperse Supramolecular Assemblies of Functional Proteins with Defined
465 Valency. *Nat. Commun.* **2015**, 6, 7134.

466 46. McCoy, K.; Douglas, T. In vivo Packaging of Protein Cargo Inside of Virus-Like Particle
467 P22. In *Virus-Derived Nanoparticles for Advanced Technologies*; Springer: 2018, pp 295-302.

468 47. Zaremba, S. M.; Gregoret, L. M., Context-dependence of Amino Acid Residue Pairing in
469 Antiparallel β-Sheets. *J. Mol. Biol.* **1999**, 291, 463-479.

470 48. Allen, M.; Poggiali, D.; Whitaker, K.; Marshall, T. R.; Kievit, R., Raincloud Plots: a
471 Multi-Platform Tool For Robust Data Visualization. *PeerJ* **2018**, 6, 27137.

472 49. Alford, R. F.; Leaver-Fay, A.; Jeliazkov, J. R.; O'Meara, M. J.; DiMaio, F. P.; Park, H.;
473 Shapovalov, M. V.; Renfrew, P. D.; Mulligan, V. K.; Kappel, K.; others, The Rosetta All-atom
474 Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **2017**, 13,
475 3031-3048.

476 50. Kuhlman, B.; Dantas, G.; Ireton, G. C.; Varani, G.; Stoddard, B. L.; Baker, D., Design of
477 a Novel Globular Protein Fold with Atomic-level Accuracy. *Science* **2003**, 302, 1364-1368.

478 51. Schymkowitz, J. W.; Rousseau, F.; Martins, I. C.; Ferkinghoff-Borg, J.; Stricher, F.;
479 Serrano, L., Prediction of Water and Metal Binding Sites and Their Affinities by Using the Fold-
480 X Force Field. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, 102, 10147-10152.

481 52. Fleishman, S. J.; Whitehead, T. A.; Ekiert, D. C.; Dreyfus, C.; Corn, J. E.; Strauch, E.-
482 M.; Wilson, I. A.; Baker, D., Computational Design of Proteins Targeting the Conserved Stem
483 Region of Influenza Hemagglutinin. *Science* **2011**, 332, 816-821.

484 53. Rocklin, G. J.; Chidyausiku, T. M.; Goreshnik, I.; Ford, A.; Houliston, S.; Lemak, A.;
485 Carter, L.; Ravichandran, R.; Mulligan, V. K.; Chevalier, A.; others, Global Analysis of Protein
486 Folding Using Massively Parallel Design, Synthesis, and Testing. *Science* **2017**, 357, 168-175.

487 54. Walsh, I.; Seno, F.; Tosatto, S. C.; Trovato, A., PASTA 2.0: an Improved Server for
488 Protein Aggregation Prediction. *Nucleic Acids Res.* **2014**, 42, 301.

489 55. Tyka, M. D.; Keedy, D. A.; André, I.; DiMaio, F.; Song, Y.; Richardson, D. C.;
490 Richardson, J. S.; Baker, D., Alternate States of Proteins Revealed by Detailed Energy
491 Landscape Mapping. *J. Mol. Biol.* **2011**, 405, 607-618.

492 56. Köker, T.; Fernandez, A.; Pinaud, F., Characterization of Split Fluorescent Protein
493 Variants and Quantitative Analyses of Their Self-Assembly Process. *Sci. Rep.* **2018**, 8, 5344.

494 57. Ito, M.; Ozawa, T.; Takada, S., Folding Coupled with Assembly in Split Green
495 Fluorescent Proteins Studied by Structure-based Molecular Simulations. *J. Phys. Chem. B* **2013**,
496 117, 13212-13218.

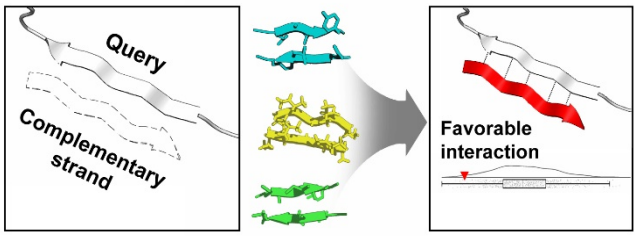
497 58. Xu, Y.; Shen, J.; Luo, X.; Zhu, W.; Chen, K.; Ma, J.; Jiang, H., Conformational
498 Transition of Amyloid β-peptide. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, 102, 5403-5407.

499 59. Miller, K. E.; Kim, Y.; Huh, W.-K.; Park, H.-O., Bimolecular Fluorescence
500 Complementation (BiFC) Analysis: Advances and Recent Applications for Genome-wide
501 Interaction Studies. *J. Mol. Biol.* **2015**, 427, 2039-2055.

502

503

504 **For Table of Contents Use Only**



505