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.

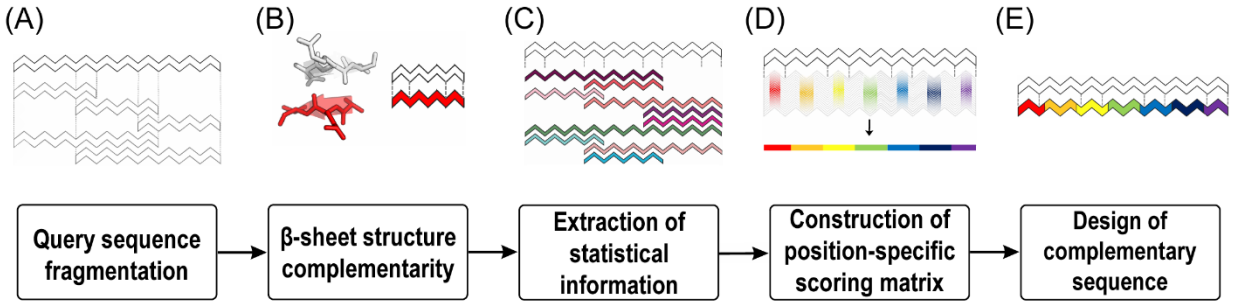
Intriguingly, significant conservation and covariations of residue pairs between neighboring β-strands have been identified in many different protein families [14]. For instance, pairs of β-branched residues and cysteines are preferred at nonhydrogen-bonded positions. Aromatic residues tend to be paired 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],

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

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

71 Herein, we present a database search method, B-SIDER (β-Stacking Interaction DEsign for
72 Reciprocity), for the design of complementary β-strands. The method generates a query-specific
73 score matrix from the structure database. To utilize the pairing information and overcome the
74 pattern noise, we hypothesized that significant complementary pairs can be amplified by
75 superposing a subset of sequence fragments. Moreover, the recent growth boom of β-sheet
76 structures [35] allows the solid statistics of β-sheet forming residue pairings [36]. Based on the
77 hypothesis and statistics of β-sheet forming residue pairings, we developed a fast and reliable
78 computational method for the design of complementary β-strand sequences. The methodology
79 augments β-sheet forming residue preferences by overlaying complementary fragment sequences
80 (**Figure 1**). We retrospectively validated our approach using a set of curated amyloidogenic targets
81 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

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

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

123 where $A_{i,p}$ is the frequency of a certain amino acid (i) at a specific complementary position (p), and
124 $O_{i,p}$ is the total count of the amino acid at p. The background frequency of a certain amino acid B_i
125 is independent of the query and thus counted from a well defined structure database(HOMSTRAD
126 database [44]). The frequency is calculated as

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

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

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

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

132

133 **Protein expression and complementation assay**

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

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

143

144 **Protein expression and purification** All constructs were transformed into BL21 (DE3)
145 Escherichia coli strains. The transformed cells were grown overnight and inoculated into a Luria-
146 Bertani media containing 50 μg/ml of kanamycin at 37 ℃. The cells were then grown until the
147 optical density of the cells reached 0.6–0.8 at 600 nm, followed by the addition of 0.7 mM of IPTG
148 (isopropyl β-D-1-thiogalactopyranoside) for induction. After incubation for 16–18 h at 18 ℃, the
149 cells were harvested and suspended in a lysis buffer containing 50 mM Tris (pH 8.0), 150 mM
150 NaCl, and 5 mM imidazole. The suspended cells were disrupted through sonication, and insoluble
151 fractions were removed using centrifugation at 18,000 g for 1 h. The supernatants were filtered
152 with 0.22 μm syringe filters and purified through affinity chromatography using Ni-NTA agarose
153 Superflow (Qiagen). The solutions were applied to the resin-packed columns and washed with a
154 buffer containing 50 mM Tris (pH 8.0), 150 mM NaCl, and 10 mM imidazole, until no proteins
155 were detected through a Bradford assay. An elution buffer (50 mM Tris (pH 7.4), 150 mM NaCl,
156 300 mM imidazole) was then applied to the column. The buffer exchange was performed with
157 phosphate buffered saline (PBS, pH 7.4) using a PD-10 column (GE health-care). The
158 concentrations of the proteins were determined by measuring the absorbance at 280 nm. All
159 purification processes were conducted at 4 ℃. The purities of the proteins were then evaluated
160 using SDS-PAGE.

161

162 **Complementation assay** The assembly of splitGFP variants was monitored and measured using
163 a fluorescence complementation assay. An excessive amount of GFP1-10 (50 pmol) in 180 µl and
164 20 µl of equal molar concentration for each GFP 11 strand (3 pmol) were co-incubated in a PBS
165 buffer (pH 7.4). The fluorescence kinetics (λ_{ex}= 488 nm /λ_{em}= 530 nm) were monitored for 12 h
166 at 25 ℃ using a TECAN infinite M200 microplate reader at 5 min intervals [37] with shaking for 2
167 sec between intervals. Each experiment was conducted in triplicate using a Nunc F 96 Micro-well
168 black plate, blocked with a solution of PBS containing 0.5 % of bovine serum albumin (BSA) for
169 30 min before the assay.

170

171 ## Results/Discussion

172 **Overview of the design process**

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

176 There are two major steps applied in the algorithm: 1) The extraction of complementarity
177 information of the β-sheet and 2) the construction of a scoring matrix (**Figure 1**). When a query
178 sequence is given, it is fragmented into several pieces of short peptides longer than three residues
179 in length, and the matching neighboring strands are collected. These fragmentation and overlaying
180 processes naturally impose weights on complementary-prone positions and amplify the pattern
181 signals. It should be noted that the subsequence search starts from the -2nd position in order to
182 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**

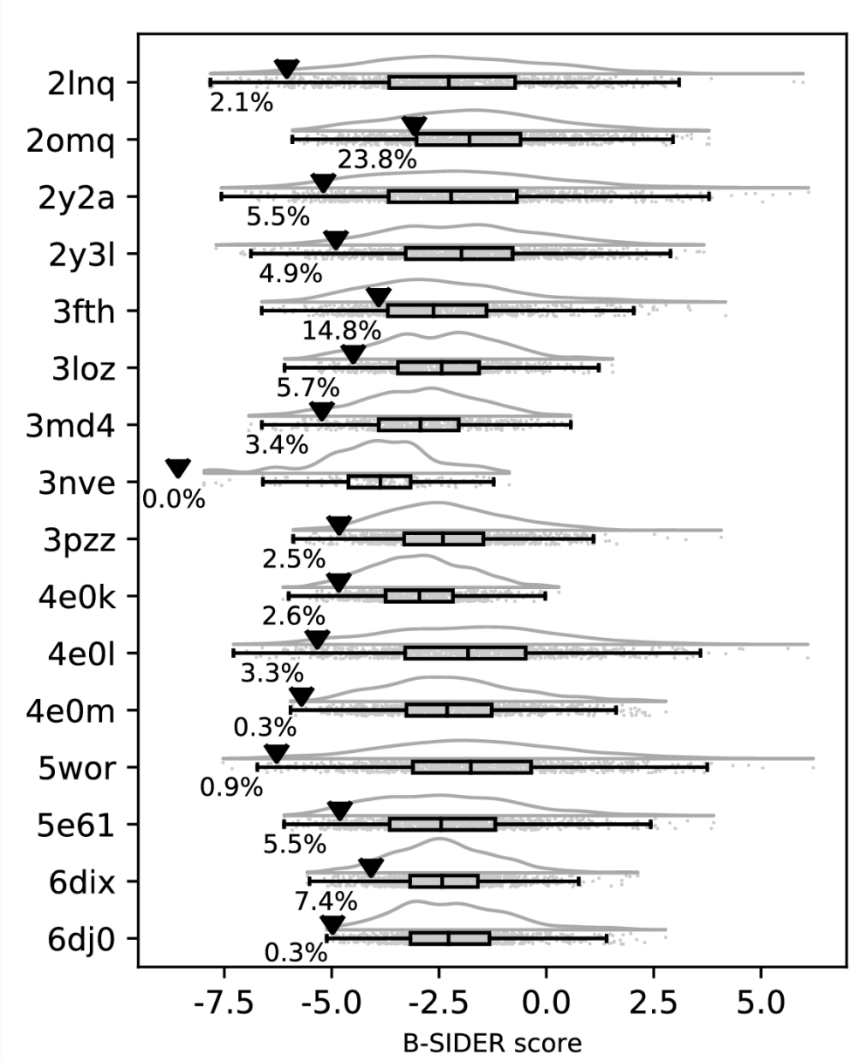
186 In an effort to validate the complementarity score, we manually curated a test set of
187 naturally occurring β-strand pairs whose environmental effects are minimal. It is known that β-
188 strand pairing is in general significantly hindered by local environments [47], whereas amyloidogenic
189 peptide 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.

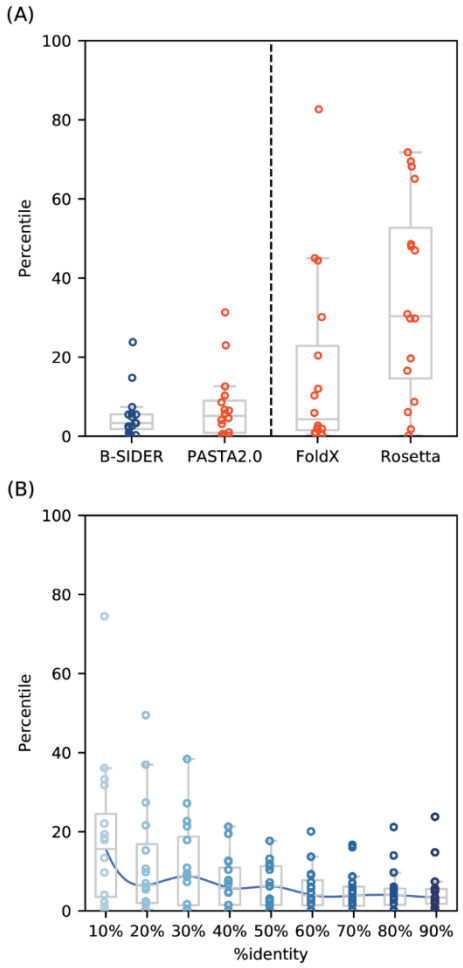
204 **Figure. 2. The predictive power of the B-SIDER score.** In this test set, native sequences were
205 compared against 1,000 random sequences. A lower score indicates more favorability. The native
206 complementarity scores are marked as (▼) and their percentile values are displayed next to the
207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

208

209 We also compared the B-SIDER score with two structure-based all-atom energy functions
210 and a sequence-based score matrix. For the structure-based methods, we picked Rosetta (Talaris
211 13) [49, 50] and FoldX [51], which have been popularly used in *de novo* protein designs [52, 53]. PASTA
212 2.0 [54] is a method for predicting regions prone to aggregation using a scoring matrix derived from
213 residue pairing patterns of β-sheets. To avoid any biases, 1,000 random sequences were newly
214 prepared per target. The "FastRelax" protocol [55] from Rosetta (Ver. 3.7), the "BuildModel"
215 command from FoldX (Ver. 4.0), and the scoring matrix from PASTA 2.0 were used against the
216 native and random sequences. The predictive power of the score function was assessed based on
217 the percentile of the native sequence score against the random sequence score distribution.

218 **Figure 3A** shows that structure-based score functions are generally worse than the
219 sequence-based scoring matrices. The results indicate that the Rosetta energy score function is not
220 sufficiently accurate for ranking complementary β-strands (35.1th percentile on average), whereas
221 the predictive powers of PASTA 2.0 and FoldX were moderate, showing 7.4th and 16.4th
222 percentiles on average respectively. B-SIDER was shown to be the most accurate in an extremely
223 consistent manner (5.2th on average. Standard deviations of Rosetta, FoldX, PASTA 2.0 and B-
224 SIDER are 25.2, 23.3, 8.7 and 6.2, respectively). Although the assessment by PASTA 2.0 is also
225 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.

236

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

255 **Prospective appplication of the algorithm to splitGFP**

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

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

267 **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

268 *Normalized values compared to the fluorescence level of the original GFP11 strand.
269

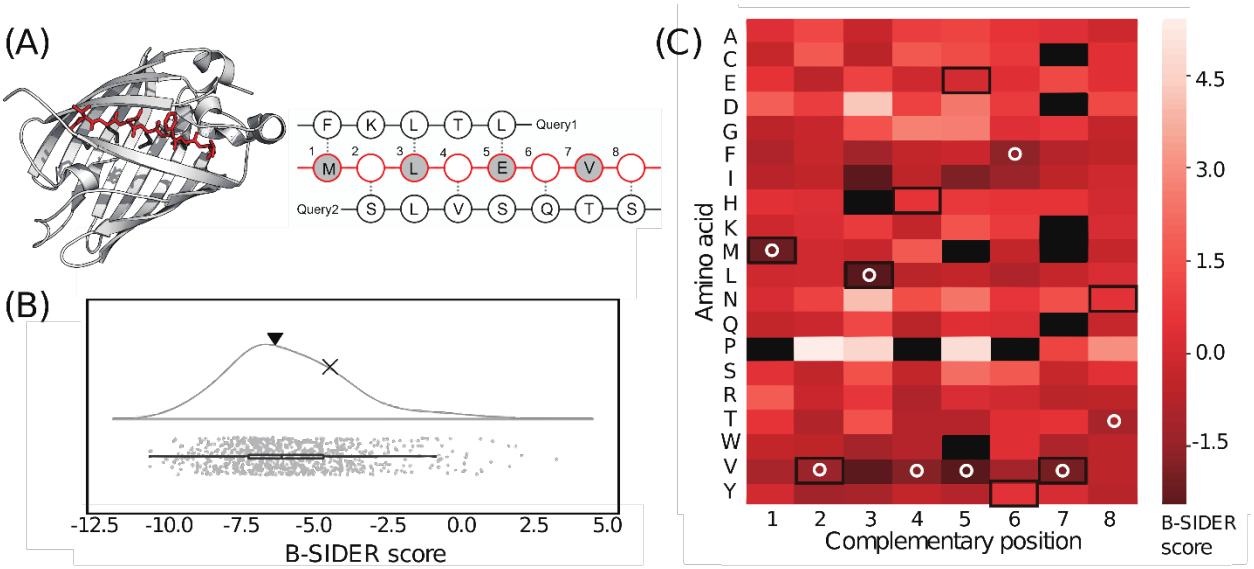
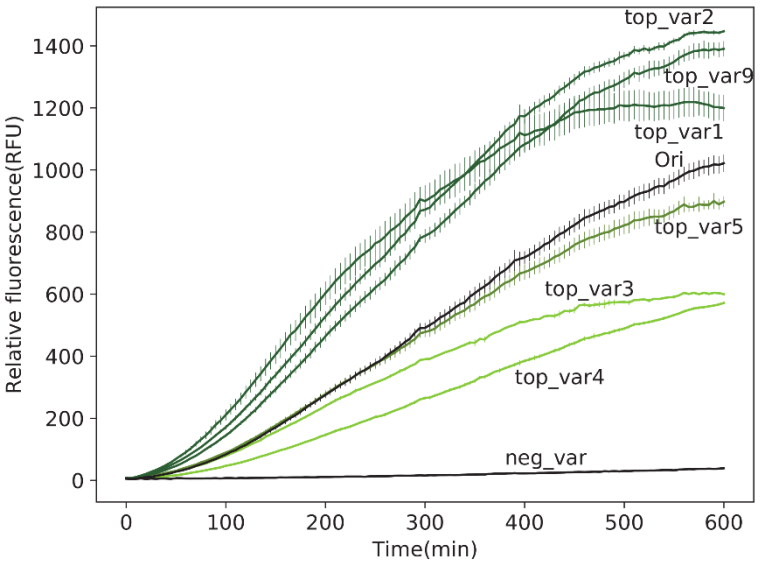


Figure. 4. Design process of GFP11 variants. (A) The strand to be designed (GFP11) is highlighted in red, and query sequences to consider are shown on the right. Dotted lines represent hydrogen bonds. The residues pointing inward, which are not subject to a mutation, are colored in gray. (B) The B-SIDER 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.



299

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

303

304 ## Conclusion

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

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

319

320 **Notes**

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

322

323 **Supporting Information**

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

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

332

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.

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

361 7. Matthes, D.; Daebel, V.; Meyenberg, K.; Riedel, D.; Heim, G.; Diederichsen, U.; Lange,
362 A.; de Groot, B. L., Spontaneous Aggregation of the Insulin-derived Steric Zipper Peptide
363 VEALYL Results in Different Aggregation Forms with Common Features. *J. Mol. Biol.* **2014**,
364 426, 362-376.

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

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

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

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

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

378 13. Zheng, W.; Tsai, M.-Y.; Chen, M.; Wolynes, P. G., Exploring the Aggregation Free
379 Energy Landscape of the Amyloid-β Protein (1-40). *Proc. Natl. Acad. Sci. U.S.A.* **2016**, 113,
380 11835-11840.

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

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

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

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

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

394 19. Bryan Jr, A. W.; Menke, M.; Cowen, L. J.; Lindquist, S. L.; Berger, B., BETASCAN:
395 Probable β-Amyloids Identified by Pairwise Probabilistic Analysis. *PLOS Comput. Biol.* **2009**, 5,
396 1000333.

397 20. Maurer-Stroh, S.; Debulpaep, M.; Kuemmerer, N.; De La Paz, M. L.; Martins, I. C.;
398 Reumers, J.; Morris, K. L.; Copland, A.; Serpell, L.; Serrano, L.; others, Exploring the Sequence
399 Determinants of Amyloid Structure Using Position-specific Scoring Matrices. *Nat. Methods*
400 **2010**, 7, 237.

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

403 22. Kim, C.; Choi, J.; Lee, S. J.; Welsh, W. J.; Yoon, S., NetCSSP: Web Application for
404 Predicting Chameleon Sequences and Amyloid Fibril Formation. *Nucleic Acids Res.* **2009**, 37,
405 469.

406 23. Tsolis, A. C.; Papandreou, N. C.; Iconomidou, V. A.; Hamodrakas, S. J., A Consensus
407 Method for the Prediction of 'Aggregation-prone' Peptides in Globular Proteins. *PLOS One* **2013**,
408 8, 54175.

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

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

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

416 27. Stranges, P. B.; Machius, M.; Miley, M. J.; Tripathy, A.; Kuhlman, B., Computational
417 Design of a Symmetric Homodimer Using β-strand Assembly. *Proc. Natl. Acad. Sci. U.S.A.*
418 **2011**, 108, 20562-20567.

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

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

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

427 31. West, M. W.; Wang, W.; Patterson, J.; Mancias, J. D.; Beasley, J. R.; Hecht, M. H., De
428 novo Amyloid Proteins from Designed Combinatorial Libraries. *Proc. Natl. Acad. Sci.* **1999**, 96,
429 11211-11216.

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

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

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

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

439 36. Sormanni, P.; Aprile, F. A.; Vendruscolo, M., Rational Design of Antibodies Targeting
440 Specific Epitopes Within Intrinsically Disordered Proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2015**,
441 112, 9902-9907.

442 37. Cabantous, S.; Terwilliger, T. C.; Waldo, G. S., Protein Tagging and Detection with
443 Engineered Self-assembling Fragments of Green Fluorescent Protein. *Nat. Biotechnol.* **2005**, 23,
444 102.

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

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

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

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

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

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

460 44. Mizuguchi, K.; Deane, C. M.; Blundell, T. L.; Overington, J. P., HOMSTRAD: a
461 Database of Protein Structure Alignments for Homologous Families. *Protein Sci.* **1998**, 7, 2469-
462 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.

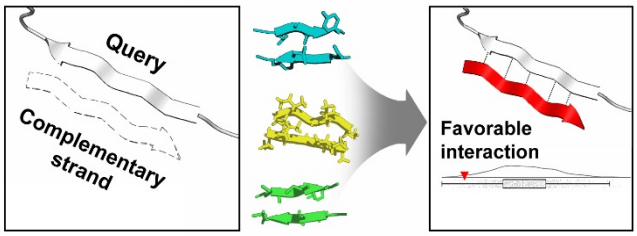
492 56. Köker, T.; Fernandez, A.; Pinaud, F., Characterization of Split Fluorescent Protein
493 Variants and Quantitative Analyses of Their Self-Assembly Process. *Sci. Rep.* **2018**, 8, 5344.

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

For Table of Contents Use Only