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

Correspondence

Yoonjoo Choi

Department of Biological Sciences, Korea Advanced Institute of Science and Technology, Daejeon, 34141, Republic of Korea.

Tel: +82-42-350-2616

Email: yoonjoo.choi@kaist.ac.kr

Hak-Sung Kim

Department of Biological Sciences, Korea Advanced Institute of Science and Technology, Daejeon, 34141, Republic of Korea.

Tel: +82-42-350-2616

Email: hskim76@kaist.ac.kr

Keywords

Beta-sheet; Beta-strand; Complementary sequence; Protein design; Amyloid; Sequence design; Protein secondary structure;

Abstract

Motivation: The β-sheet is an element of protein secondary structure, and intra- and inter-molecular β-sheet interactions play pivotal roles in biological regulatory processes including scaffolding, transporting, and oligomerization. In nature, β-sheet formation is tightly regulated, because dysregulated β-stacking often leads to severe diseases such as Alzheimer's, Parkinson's, systemic amyloidosis and diabetes. Thus, the identification of intrinsic β-sheet forming propensities could provide valuable insight into protein design for the development of novel therapeutics. However, structure-based design methods may not be generally applicable to such amyloidogenic peptides mainly due to high structural plasticity and complexity. Therefore, an alternative design strategy based on complementary sequence information is of great significance.

Results: We developed B-SIDER (β-Sheet Interaction DEsign for Reciprocity), a database search method for the design of complementary β-strands. The method makes use of the structural database information and generates a query-specific score matrix. 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 was able to distinguish wild-type amyloidogenic β-strands as favored interactions in a more consistent manner than the other methods. B-SIDER is then prospectively applied to the design of complementary β-strands for the splitGFP scaffold. Three variants were identified to have stronger interactions than its original sequence selected by directed evolution, emitting higher fluorescence intensities. Our results support that B-SIDER can be applicable to the design of other β-strands, assisting in the development of therapeutics against disease-related amyloidogenic peptides.

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

Introduction

The β-sheet is one of the major units of protein structure (Bhattacharjee and Biswas, 2010), and plays a variety of functional roles in transportation, recognition, scaffolding and enzymatic processes (Marcos, et al., 2018). Recently, the mechanism of β-sheet formation has received much attention because of its close relations with several critical diseases such as Alzheimer's disease, Parkinson's disease, type 2 diabetes and systemic amyloidosis (Chiti and Dobson, 2017; Richardson and Richardson, 2002). Such diseases are known to be linked to the precipitation of dysregulated β-stacking between neighboring β-strands (Colletier, et al., 2011; Liu, et al., 2012; Matthes, et al., 2014). In this regard, the information about 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 (Giorgetti, et al., 2018).

While structure-based protein design approaches have shown notable successes in several cases (Huang, et al., 2016), their application to *de novo* β-sheet designs still remains challenging (Dou, et al., 2018; Marcos, et al., 2018). Structure-based design approaches require a well-defined protein structure, but amyloidogenic peptides usually have highly disordered structures (Jang, et al., 2016). Structural identification of such peptides has long been hindered by high degrees of structural plasticity, transiency, and complexity due to self-oligomerization (Dovidchenko and Galzitskaya, 2015; Zheng, et al., 2016). 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 were identified in many protein families (Mandel-Gutfreund, et al., 2001). For instance, pairs of β-branched residues and cysteines are preferred at nonhydrogen-bonded positions.

74 Aromatic residues tend to be paired with valine or glycine (Steward and Thornton, 2002). Several
75 computational algorithms have been developed to predict aggregation-prone regions based on the
76 internal β-sheet forming patterns. While different in details, they make use of either statistical
77 potentials such as Tango (Fernandez-Escamilla, et al., 2004), PASTA (Trovato, et al., 2006),
78 SALSA (Zibaee, et al., 2007), BETASCAN (Bryan Jr, et al., 2009), and Waltz (Maurer-Stroh, et
79 al., 2010) or physicochemical properties of amino acids (Tartaglia and Vendruscolo, 2008).
80 Additionally, consensus methods and machine-learning approaches have also been developed
81 (Kim, et al., 2009; Tsolis, et al., 2013), showing fine agreements with experimental results.

82 It has been reported that β-strand interactions can be stabilized by introducing the β-sheet
83 favored pairs (Kortemme, et al., 1998; Minor Jr and Kim, 1994; Quinn, et al., 1994; Stranges, et
84 al., 2011) and charge pairing between neighboring β-strands (Shammas, et al., 2011; Wang and
85 Hecht, 2002; West, et al., 1999). Recent studies showed that fragments derived from the
86 amyloidogenic region can be used for β-stacking modeling (Gallardo, et al., 2016; Liu, et al., 2012).
87 While the use of the amino acid pairing information in protein design has been attempted in
88 elsewhere, practical applications of such patterns have been limited mainly owing to the lack of
89 comprehensive quantification for residue pairing and noisy patterns of β-sheet forming residue
90 pairs (Bhattacharjee and Biswas, 2010; Fujiwara, et al., 2012; Hutchinson, et al., 1998). The β-
91 sheet forming peptides appear to have poor sequence commonalities and imperfect repeats (Bryan
92 Jr, et al., 2009). Therefore, careful curation of meaningful patterns is required for the practical
93 protein design strategy of complementary β-strands.

94 Herein, we present a database search method, B-SIDER (β-Stacking Interaction DEsign for
95 Reciprocity), to design complementary β-strands. The method generates a query-specific score
96 matrix from the structure database. To utilize the pairing information and overcome the pattern

97 noise, we hypothesized that significant complementary pairs can be amplified by superposing a
98 subset of sequence fragments. Moreover, the recent growth boom of β-sheet structures (Marcos
99 and Silva, 2018) allows the solid statistics of β-sheet forming residue pairings (Sormanni, et al.,
100 2015). Based on the hypothesis and statistics of β-sheet forming residue pairings, we developed a
101 fast and reliable computational method for the design of complementary β-strand sequences. The
102 methodology augments β-sheet forming residue preferences through overlaying complementary
103 fragment sequences (**Fig. 1**). We retrospectively validated our approach using a set of curated
104 amyloidogenic targets and compared it with two popularly used structure-based methods (Rosetta
105 and FoldX) and a sequence-based aggregation prediction algorithm (PASTA2.0). Our algorithm
106 was shown to clearly distinguish favorable β-sheet forming sequences entirely based on the query
107 sequence, whereas the structure-based energy functions exhibited inconsistent results depending
108 on targets. The utility and potential of our method were demonstrated by designing novel
109 complementary peptides for splitGFP. The designed sequences showed stronger interactions with
110 neighboring strands of the scaffold and consequently higher fluorescence emissions than the
111 original peptide selected by directed mutagenesis (Cabantous, et al., 2005).

112

113 ## Methods

114 **Computational algorithm for the design of complementary sequences**

115 ***Collection of β-strand information***

116 Non-redundant structures determined by high resolution X-ray crystallography were
117 collected from the PDB: < 90 % sequence identity, < 3 Å resolution. Given the query target
118 sequence, the non-redundant structure database was used to extract pairing information from

119 matched sequences. Initially, the target sequence is divided into linear moving-windows whose
120 residues in length range from 3 to the entire target sequence length. Any structures with identical
121 target sequence fragments to the split queries were collected, followed by further filtering based
122 on the definition of β-sheet secondary structure (the distance between backbone nitrogen-oxygen
123 atom pairs < 5 Å). In order to remove redundancy, protein structures that contain the matches were
124 compared using TMalign (Zhang and Skolnick, 2005). If TM-score > 0.7 and sequence identity >
125 90 %, one of the matched sequences was removed.

126 While the method is applicable to both parallel and anti-parallel β-sheets in theory, we
127 mainly focused on anti-parallel β-sheets in this study since anti-parallel cases are more frequently
128 observed compared to parallel ones (Hubbard, 1994). Disease-related amyloidogenesis is also
129 known to be initiated with anti-parallel β-sheets and soluble oligomeric amyloid species mainly
130 exist as anti-parallel (Cerf, et al., 2009; Gordon, et al., 2004).

131

132 ***Complementary sequence score***

133 The β-sheet complementarity score function is derived from the environment-specific
134 substitution score (Choi and Deane, 2010). We hypothesized that each position of a β-strand is
135 independent of one another, and their complementarities are determined by residue pairs from
136 neighboring strands. Given the query sequence, all of the identified neighboring sequences are
137 pooled together as described in the previous section. The amino acid frequency at each
138 complementary position is counted as

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

139 where $A_{i,p}$ is the frequency of a certain amino acid i at a specific complementary position p. $O_{i,p}$ is

140 the total count of the amino acid at p. The background frequency of the certain amino acid (B_i) is

141 counted from the HOMSTRAD database (Mizuguchi, et al., 1998) and calculated as

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

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

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

143 It should be noted that the complementarity score is completely data-driven, i.e. if an amino

144 acid never appears at a certain position, a high penalty score is imposed. We only consider

145 complementary amino acids which are found at least once in the entire identified sequences. The

146 final score is the sum of the scores at all of the positions.

147

148 **Protein expression and complementation assay**

149 ***Gene construction***

150 The gene coding for splitGFP (Cabantous, et al., 2005) consists of the 1-10th strands (GFP1-

151 10) and 11th strand (GFP11) template (**Table S1**). They were cloned into pET-28a (Novagen)

152 vector between the Nde-I and Xho-I restriction sites. We introduced additional mutations to GFP1-

153 10 to inhibit aggregation and convenient expression (Kim, et al., 2015). The GFP11 strand was

154 fused with a P22 virus-like particle scaffolding protein (McCoy and Douglas, 2018) for soluble

155 and stable expression. Mutations on GFP11 were introduced by PCR using the mutagenic primers

156 (**Table S2**), and the resulting genes were cloned into the pET-28a vector. Six histidine residues

157 were fused to the N-terminal of the GFP1-10 and GFP11 genes as an affinity purification tag.

158

159 ***Protein expression and purification***

160 All of the constructs were transformed into BL21 (DE3) *E. coli* strains. The transformed
161 cells were grown overnight and inoculated into a Luria-Bertani media containing 50 μg/ml of
162 kanamycin at 37 ℃. Then, cells were grown until an optical density of the cells reached 0.6-0.8 at
163 600 nm, followed by addition of 0.7 mM of IPTG (isopropyl β-D-1-thiogalactopyranoside) for
164 induction. After incubation for 16-18 hours at 18 ℃, the cells were harvested and suspended in a
165 lysis buffer containing 50 mM Tris (pH 8.0), 150 mM NaCl, and 5 mM imidazole. The suspended
166 cells were disrupted by sonication, and insoluble fractions were removed by centrifugation at
167 18,000 g for 1 hour. The supernatants were filtered using 0.22 μm syringe filters and purified
168 through affinity chromatography with Ni-NTA agarose Superflow (Qiagen). The solutions were
169 applied to the resin-packed columns and washed with a buffer containing 50 mM Tris (pH 8.0),
170 150 mM NaCl, and 10 mM imidazole, until no protein was detected by Bradford assay. Then, an
171 elution buffer (50 mM Tris (pH 7.4), 150 mM NaCl, 300 mM imidazole) was applied to the column.
172 The buffer exchange was performed by PD-10 column (GE health-care) to PBS (phosphate
173 buffered saline, pH 7.4). The concentrations of the proteins were determined by measuring the
174 absorbance at 280 nm. All the purification processes were performed at 4 ℃. The purities of
175 proteins were then evaluated by SDS-PAGE.

176

177 ***Complementation assay***

178 The assembly of splitGFP variants was monitored and measured by fluorescence
179 complementation assay. Excessive amount of GFP1-10 (50 pmol) in 180 μl and 20 μl of equal

180 molar concentration of each GFP 11 strand (3 pmol) were co-incubated in PBS buffer (pH 7.4).
181 Fluorescence kinetics (λ_{ex}= 488 nm /λ_{em}= 530 nm) were monitored for 12 hours at 25 ℃ by
182 TECAN infinite M200 microplate reader at 5 minutes intervals (Cabantous, et al., 2005) with
183 shaking for 2 seconds between intervals. Each experiment was performed in triplicate with Nunc
184 F 96 Micro-well black plate, blocked with a solution of PBS containing 0.5 % of Bovine serum
185 albumin (BSA) for 30 minutes before the assay.

186

187 ## Results and Discussion

188 **Overview of the design process**

189 We hypothesized that repetitively observed amino acid pairing patterns indicate the
190 "smoking-gun" of strong preferences to β-sheet. It was also assumed that the sequence with most
191 frequent patterns would directly form a β-sheet without considering other environmental
192 contributions.

193 There are two major steps in the algorithm: 1) The extraction of β-sheet complementarity
194 information and 2) the construction of scoring matrix (See **Fig. 1** and **Fig. S1**). When a query
195 sequence is given, it is fragmented into several pieces of short peptides longer than three residues
196 in length and matched neighboring strands are collected. This fragmentation and overlaying
197 processes naturally impose weights on complementary-prone positions and amplify pattern signals
198 (**Fig. S1**). After the collection of matched sequences, a position-specific complementarity scoring
199 matrix is constructed. The obtained scoring matrix is used to evaluate and design the
200 complementarity of β-strand interactions.

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

202 In an effort to validate the complementarity score, we manually curated a test set of
203 naturally occurring β-strand pairs whose environmental effects are minimal. It is known that β-
204 strand pairing is in general greatly hindered by local environments (Zaremba and Gregoret, 1999),
205 but amyloidogenic peptide segments are known to form natural β-sheets in themselves (Trovato,
206 et al., 2006). We thus selected a set of widely known amyloidogenic structures whose aggregation-
207 prone regions have been identified (**Table 1** and **Table S3**).

208 To assess the complementarity of the native sequences, we compared their scores with
209 those of random sequences. The natural amyloidogenic segments are known to be highly
210 aggregation-prone, so they are expected to be highly preferred, i.e., having fairly low scores in the
211 random sequence score distributions. **Figure 2** shows that all the native sequence scores are ranked
212 extremely low in all of the distributions. On average, the native sequences are within 4.1 % of the
213 distributions (**Fig. 2**). The results indicate that the scoring function is extremely useful in detecting
214 favorable β-strand counterparts.

215 We also compared the B-SIDER score with two structure-based all-atom energy functions
216 and a sequence-based score matrix. For structure-based methods, we picked Rosetta (Talaris 13)
217 (Alford, et al., 2017; Kuhlman, et al., 2003) and FoldX (Schymkowitz, et al., 2005), which have
218 been popularly used in *de novo* protein designs (Fleishman, et al., 2011; Rocklin, et al., 2017).
219 PASTA2.0 (Walsh, et al., 2014) is a method to predict aggregation-prone regions using the scoring
220 matrix derived from residue pairing patterns of β-sheets. To avoid any biases, 1,000 random
221 sequences were newly prepared per target. "FastRelax" protocol (Tyka, et al., 2011) from Rosetta
222 (Ver. 3.7), "BuildModel" command from FoldX (Ver. 4.0) and the scoring matrix from PASTA2.0
223 were used against the native and the random sequences. The predictive power of a score function

224 was assessed by the percentile value of the native sequence score against the random sequence
225 score distribution.

226 **Figure 3** shows that structure-based score functions are in general worse than the sequence-
227 based scoring matrices. The results indicate that the Rosetta energy score function is not
228 sufficiently accurate for ranking complementary β-strands (35.8th percentile on average), whereas
229 the predictive powers of PASTA2.0 and FoldX were moderate, showing 10.8th and 14.7th,
230 respectively. B-SIDER was shown to be the most accurate in an extremely consistent manner.
231 While the assessment of PASTA2.0 is also fairly consistent, the query-specific nature of B-SIDER
232 may give better results.

233 Considering the Rosetta relax protocol performs flexible backbone refinements, the use of
234 the fixed-backbone calculation seems to be better for the evaluation of β-sheet complementarity.
235 It should be noted that the inconsistent results of Rosetta imply that FoldX prediction would be
236 also highly driven by structure preparation, i.e. design with ill-defined models may not be generally
237 successful. On the other hand, B-SIDER and PASTA2.0 do not depend on query structures, and
238 thus, it can be applied to general cases such as β-sheet interactions with high structural plasticity
239 and poor structural integrity, which are the common features of amyloidogenic peptides.
240 Furthermore, the process of collecting complementary motifs of B-SIDER also appeared to be
241 powerful, making it possible to distinguish favorable complementary sequences not easily detected
242 by one-to-one residue pairing.

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

244 As shown in the retrospective test, B-SIDER is extremely useful in discriminating naturally
245 β-strand forming sequences. As a proof of concept, we prospectively designed novel

246 complementary β-strands for splitGFP. SplitGFP is a fragmented protein pair derived from
247 superfolderGFP (Cabantous, et al., 2005), comprising a scaffold containing 10 β-strands (GFP1-
248 10) and its complementary β-strand peptide (GFP11). GFP11 specifically interacts with GFP1-10
249 and the strand tightly forms a stable β-sheet structure, which facilitates the chromophore
250 maturation in an irreversible manner (Köker, et al., 2018). This assembly process results in the
251 emission of the green fluorescence. Because GFP11 is known to be disordered in solution, its
252 conformational transition from the disordered to induced β-sheet is similar to amyloidogenic
253 peptides (Ito, et al., 2013; Xu, et al., 2005). This model system thus efficiently assesses whether
254 designed sequences by B-SIDER have favorable β-sheet interactions.

255 Original GFP11 was designed by directed mutagenesis, and it shows a high intrinsic
256 propensity to form hydrogen bonds with the neighboring β-strands of GFP1-10 (Miller, et al.,
257 2015). In our case, the queries are the neighboring strands of GFP1-10 (**Fig. 4A**). It is known that
258 the residues pointing inward (1, 3, 5, and 7th positions) directly interact with the chromophore and
259 thus they were not subject to mutation. B-SIDER identified 2,637 non-redundant sequences from
260 the structure database. The native sequence is ranked at a modest score among randomly chosen
261 1,000 possible sequence variants (46th percentile), indicating that there could be room for
262 complementary sequences with stronger interactions than the original one (**Fig. 4B**). We then
263 selected 10 sequences with the lowest B-SIDER scores (top_vars; **Table 2**). Amino acid
264 compositions of the 10 variants are mostly hydrophobic or branched amino acids (**Fig. S2**).
265 Additionally, one sequence with a high score (> 75th percentile) was randomly selected as a
266 negative control (neg_var, 77th).

267 The selected variants were successfully expressed and purified (**Fig.S3**) except for four
268 clones (top_var6, 7, 8, and 10) which were observed to be insoluble, perhaps due to aggregation.

269 Among those expressed, three variants (top_var1, 2, and 9) showed faster assembly patterns and
270 higher signals compared to the original GFP11 (**Fig. 5**). No functional aberrance with excitation
271 and emission was observed (**Fig. S4**). All the successful variants, which emitted stronger
272 fluorescence levels than the original one, were shown to have the pair of phenylalanine and
273 threonine at the positions 6 and 8, respectively. These results demonstrate that the designed
274 variants indeed formed complementary β-strands in a more favorable fashion than its original
275 peptide as predicted. The other variants showed slightly lower signals than the original one, but
276 still gave rise to clear fluorescence signals (**Fig. 5**). The negative control (neg_var) barely emitted
277 any signal, suggesting that the score indeed indicates the complementarity of β-stacking
278 interactions. We also assigned scores of the GFP11 variants using other scoring methods. As
279 shown in the retrospective test set, Rosetta and FoldX were not able to discriminate top_vars as
280 favorable (**Fig. S6**). However, PASTA2.0 was again fairly accurate in this case.

281

282 **Conclusion**

283 β-sheet forming patterns are crucial for understanding the aggregation mechanism of
284 disease-related β-sheets and developing potential therapeutics against them. Unlike α-helices,
285 however, there has been no established design principle for the complementarity of β-sheets. In
286 this study, we developed B-SIDER, a database search method for the design of complementary β-
287 strands based on the intrinsic β-sheet forming propensities. Statistical patterns of interacting
288 residue pairs between neighboring β-strands enable to quantify the complementary interaction. We
289 demonstrated that the statistical potential can be directly applied to the design of complementary
290 β-strand sequences. Using splitGFP as a model system, we successfully designed fragment variants,
291 which led to stronger fluorescence emissions than the native one originally identified by 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 to protein designs where structure-based methods are not effective, including the development of protein binders specifically against disease-related intrinsically disordered proteins.

Acknowledgements

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

Funding

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-2017R1A2A 1A05001091) of the National Research Foundation (NRF) funded by the Ministry of Science and ICT.

References

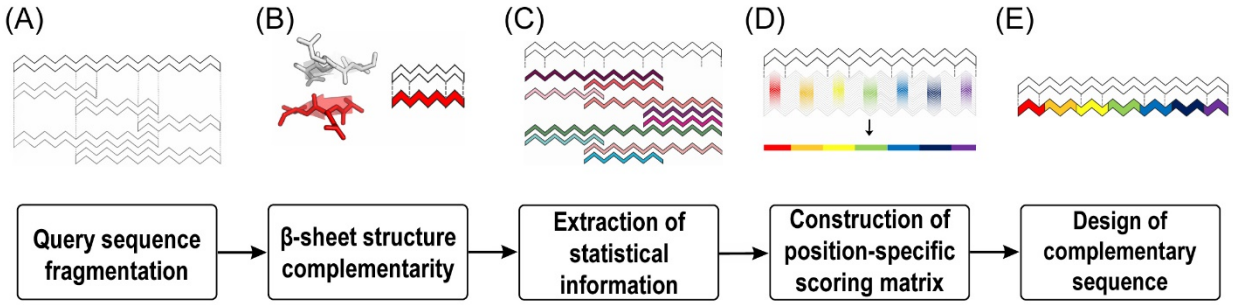
1. Alford, R.F., *et al.* The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* 2017;13(6):3031-3048.
2. Allen, M., *et al.* Raincloud plots: a multi-platform tool for robust data visualization. *PeerJ* 2018;6:27137.
3. Bhattacharjee, N. and Biswas, P. Position-specific propensities of amino acids in the β-strand. *BMC Struct. Biol.* 2010;10(1):29.
4. Bryan Jr, A.W., *et al.* BETASCAN: probable β-amyloids identified by pairwise probabilistic analysis. *PLOS Comput. Biol.* 2009;5(3):1000333.
5. Cabantous, S., Terwilliger, T.C. and Waldo, G.S. Protein tagging and detection with engineered self-assembling fragments of green fluorescent protein. *Nat. Biotechnol.* 2005;23(1):102.

6. Cerf, E., *et al.* Antiparallel β-sheet: a signature structure of the oligomeric amyloid-β peptide. *Biochem. J.* 2009;421(3):415-423.
7. Chiti, F. and 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.
8. Choi, Y. and Deane, C.M. FREAD revisited: accurate loop structure prediction using a database search algorithm. *Proteins* 2010;78(6):1431-1440.
9. Colletier, J.-P., *et al.* Molecular basis for amyloid-β polymorphism. *Proc. Natl. Acad. Sci. U.S.A.* 2011;108(41):16938-16943.
10. Dou, J., *et al.* De novo design of a fluorescence-activating β-barrel. *Nature* 2018;561(7724):485.
11. Dovidchenko, N.V. and Galzitskaya, O.V. Computational approaches to identification of aggregation sites and the mechanism of amyloid growth. In, *Lipids in Protein Misfolding*. Springer; 2015. p. 213-239.
12. Fernandez-Escamilla, A.-M., *et al.* Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.* 2004;22(10):1302.
13. Fleishman, S.J., *et al.* Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science* 2011;332(6031):816-821.
14. Fujiwara, K., Toda, H. and Ikeguchi, M. Dependence of α-helical and β-sheet amino acid propensities on the overall protein fold type. *BMC Struct. Biol.* 2012;12(1):18.
15. Gallardo, R., *et al.* De novo design of a biologically active amyloid. *Science* 2016;354(6313):4949.
16. Giorgetti, S., *et al.* Targeting amyloid aggregation: an overview of strategies and mechanisms. *Int. J. Mol. Sci.* 2018;19(9):2677.
17. Gordon, D.J., *et al.* Increasing the amphiphilicity of an amyloidogenic peptide changes the β-sheet structure in the fibrils from antiparallel to parallel. *Biophys. J.* 2004;86(1):428-434.
18. Huang, P.-S., Boyken, S.E. and Baker, D. The coming of age of de novo protein design. *Nature* 2016;537(7620):320.
19. 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. p. 336-336.
20. Hutchinson, E.G., *et al.* Determinants of strand register in antiparallel β-sheets of proteins. *Protein Sci.* 1998;7(11):2287-2300.
21. Ito, M., Ozawa, T. and Takada, S. Folding Coupled with Assembly in Split Green Fluorescent Proteins Studied by Structure-based Molecular Simulations. *J. Phys. Chem. B* 2013;117(42):13212-13218.
22. Jang, H., *et al.* Computational methods for structural and functional studies of Alzheimer's amyloid ion channels. In, *Protein Amyloid Aggregation*. Springer; 2016. p. 251-268.
23. Kim, C., *et al.* NetCSSP: web application for predicting chameleon sequences and amyloid fibril formation. *Nucleic Acids Res.* 2009;37(suppl_2):469.
24. Kim, Y.E., *et al.* Green fluorescent protein nanopolygons as monodisperse supramolecular assemblies of functional proteins with defined valency. *Nat. Commun.* 2015;6:7134.
25. Köker, T., Fernandez, A. and Pinaud, F. Characterization of Split Fluorescent Protein Variants and Quantitative Analyses of Their Self-Assembly Process. *Sci. Rep.* 2018;8(1):5344.

26. Kortemme, T., Ramírez-Alvarado, M. and Serrano, L. Design of a 20-amino acid, three-stranded β-sheet protein. *Science* 1998;281(5374):253-256.
27. Kuhlman, B.*, et al.* Design of a novel globular protein fold with atomic-level accuracy. *Science* 2003;302(5649):1364-1368.
28. Liu, C.*, et al.* Out-of-register β-sheets suggest a pathway to toxic amyloid aggregates. *Proc. Natl. Acad. Sci.* 2012;109(51):20913-20918.
29. Mandel-Gutfreund, Y., Zaremba, S.M. and 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(5):1145-1159.
30. Marcos, E.*, et al.* De novo design of a non-local β-sheet protein with high stability and accuracy. *Nat. Struct. Mol. Biol.* 2018;25(11):1028.
31. Marcos, E. and Silva, D.-A. Essentials of de novo protein design: Methods and applications. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 2018;8(6):1374.
32. Matthes, D.*, et al.* Spontaneous aggregation of the insulin-derived steric zipper peptide VEALYL results in different aggregation forms with common features. *J. Mol. Biol.* 2014;426(2):362-376.
33. Maurer-Stroh, S.*, et al.* Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat. Methods* 2010;7(3):237.
34. McCoy, K. and Douglas, T. In Vivo Packaging of Protein Cargo Inside of Virus-Like Particle P22. In, *Virus-Derived Nanoparticles for Advanced Technologies*. Springer; 2018. p. 295-302.
35. Miller, K.E.*, et al.* Bimolecular fluorescence complementation (BiFC) analysis: advances and recent applications for genome-wide interaction studies. *J. Mol. Biol.* 2015;427(11):2039-2055.
36. Minor Jr, D.L. and Kim, P.S. Measurement of the β-sheet-forming propensities of amino acids. *Nature* 1994;367(6464):660.
37. Mizuguchi, K.*, et al.* HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.* 1998;7(11):2469-2471.
38. Quinn, T.P.*, et al.* Betadoublet: de novo design, synthesis, and characterization of a beta-sandwich protein. *Proc. Natl. Acad. Sci.* 1994;91(19):8747-8751.
39. Richardson, J.S. and 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(5):2754-2759.
40. Rocklin, G.J.*, et al.* Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* 2017;357(6347):168-175.
41. Schymkowitz, J.W.*, et al.* Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proc. Natl. Acad. Sci.* 2005;102(29):10147-10152.
42. Shammas, S.L.*, et al.* Perturbation of the stability of amyloid fibrils through alteration of electrostatic interactions. *Biophys. J.* 2011;100(11):2783-2791.
43. Sormanni, P., Aprile, F.A. and Vendruscolo, M. Rational design of antibodies targeting specific epitopes within intrinsically disordered proteins. *Proc. Natl. Acad. Sci.* 2015;112(32):9902-9907.
44. Steward, R.E. and Thornton, J.M. Prediction of strand pairing in antiparallel and parallel β-sheets using information theory. *Proteins* 2002;48(2):178-191.
45. Stranges, P.B.*, et al.* Computational design of a symmetric homodimer using β-strand assembly. *Proc. Natl. Acad. Sci.* 2011;108(51):20562-20567.

46. Tartaglia, G.G. and Vendruscolo, M. The Zyggregator method for predicting protein aggregation propensities. *Chem. Soc. Rev.* 2008;37(7):1395-1401.
47. Trovato, A.*, et al.* Insight into the structure of amyloid fibrils from the analysis of globular proteins. *PLOS Comput. Biol.* 2006;2(12):170.
48. Tsolis, A.C.*, et al.* A consensus method for the prediction of 'aggregation-prone' peptides in globular proteins. *PLoS One* 2013;8(1):54175.
49. Tyka, M.D.*, et al.* Alternate states of proteins revealed by detailed energy landscape mapping. *J. Mol. Biol.* 2011;405(2):607-618.
50. Walsh, I.*, et al.* PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Res.* 2014;42(W1):301.
51. Wang, W. and Hecht, M.H. Rationally designed mutations convert de novo amyloid-like fibrils into monomeric β-sheet proteins. *Proc. Natl. Acad. Sci.* 2002;99(5):2760-2765.
52. West, M.W.*, et al.* De novo amyloid proteins from designed combinatorial libraries. *Proc. Natl. Acad. Sci.* 1999;96(20):11211-11216.
53. Xu, Y.*, et al.* Conformational transition of amyloid β-peptide. *Proc. Natl. Acad. Sci.* 2005;102(15):5403-5407.
54. Zaremba, S.M. and Gregoret, L.M. Context-dependence of Amino Acid Residue Pairing in Antiparallel β-Sheets. *J. Mol. Biol.* 1999;291(2):463-479.
55. Zhang, Y. and Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* 2005;33(7):2302-2309.
56. Zheng, W.*, et al.* Exploring the aggregation free energy landscape of the amyloid-β protein (1-40). *Proc. Natl. Acad. Sci.* 2016;113(42):11835-11840.
57. Zibaee, S.*, et al.* 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(5):906-918.

436 **Fig. 1. Overview of the B-SIDER algorithm** (A) When a query sequence is given, the query is
437 divided into smaller linear peptides ranging from 3 residues to its full length, and exactly matched
438 sequences are identified. (B) The matched sequences are checked against the structure database so
439 that the matches indeed form β-sheets. (C) If the matches form β-sheets, their complementary
440 sequences are extracted. (D) A position-specific score matrix is constructed based on the
441 complementary sequence information. (E) Final complementary sequences are designed using the
442 score matrix.



443

Fig. 2. The predictive power of the B-SIDER score. In this test set, native sequences were compared against 1,000 random sequences. The lower, the more favorable. 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 (Allen, et al., 2018)

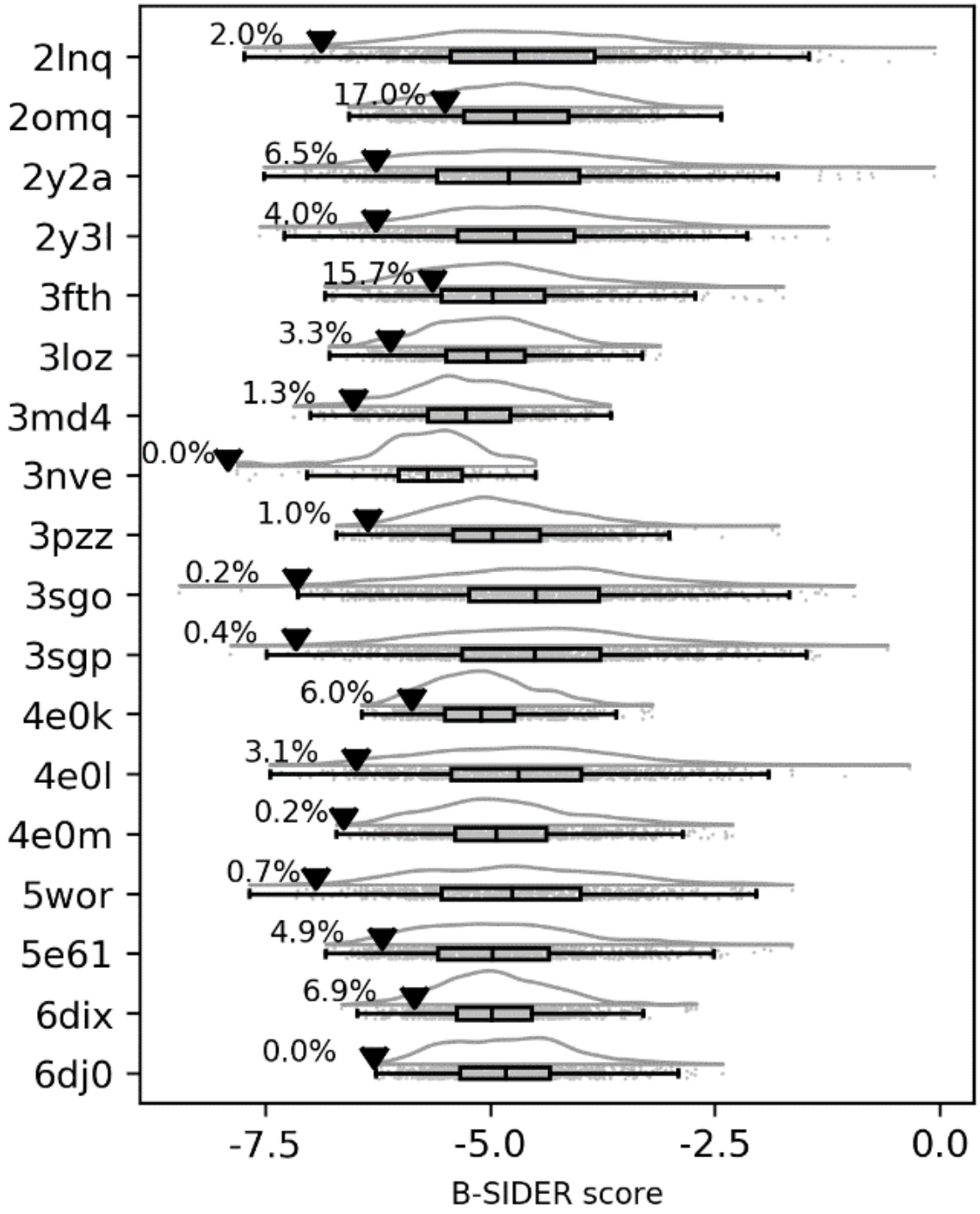


Fig. 3. The comparison of the B-SIDER score with the structure-based energy scores and existing sequence-based method. The B-SIDER score is compared against two popularly used structure-based protein design methods (Rosetta and FoldX) and the scroring matrix for the prediction of amyloid formation (PASTA2.0). For a fair comparison, we generated new sets of 1,000 random sequences per target and assessed their scores using each energy function. (A) The B-SIDER score gave the most consistent assessment. (B) The percentile values of native complementary sequences against the score distributions of the randomly generated sequences are presented. Overall, the sequence-based approaches showed higher accuracies. The better performance of FoldX than Rosetta may indicate that the fixed-backbone design strategy may be more useful in the design and prediction of β-sheet complementarity.

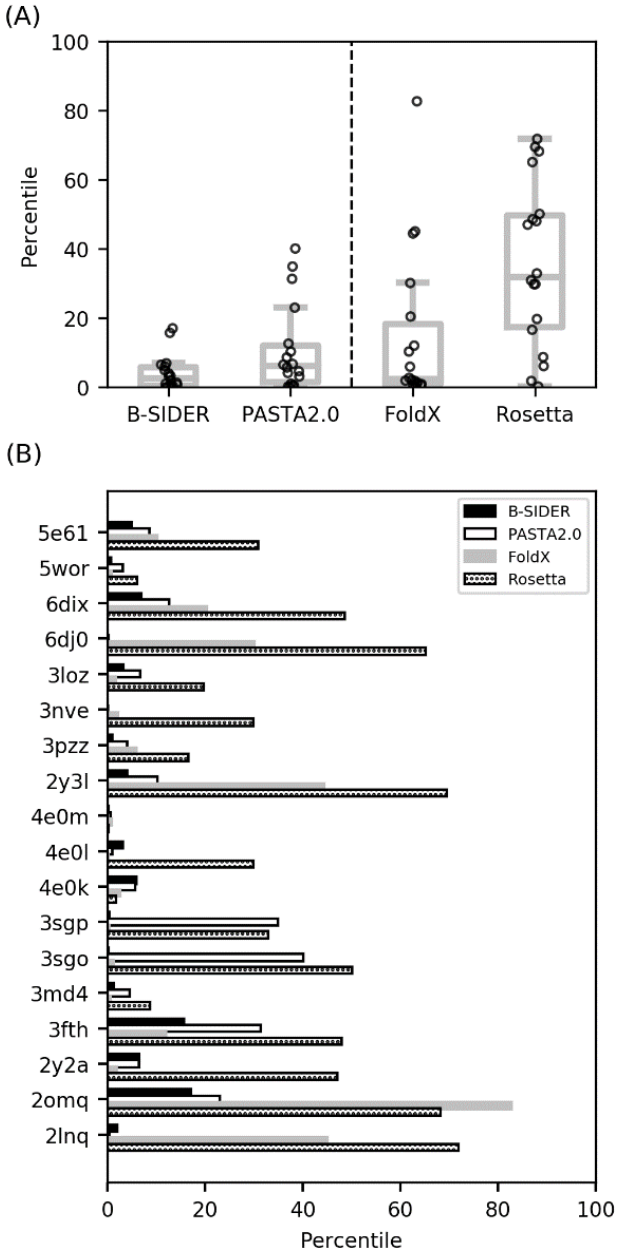
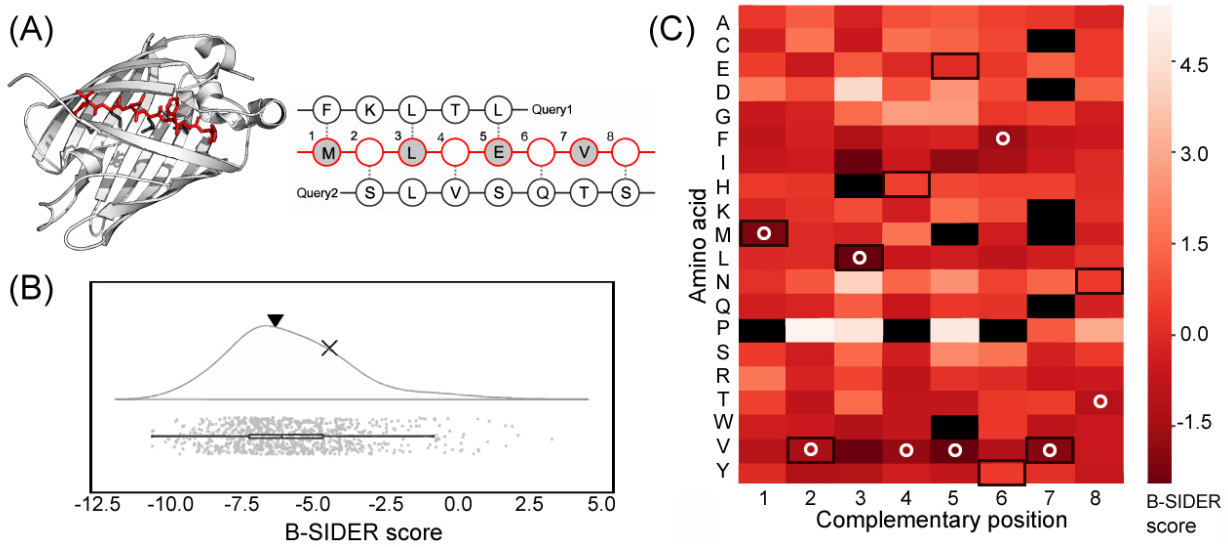
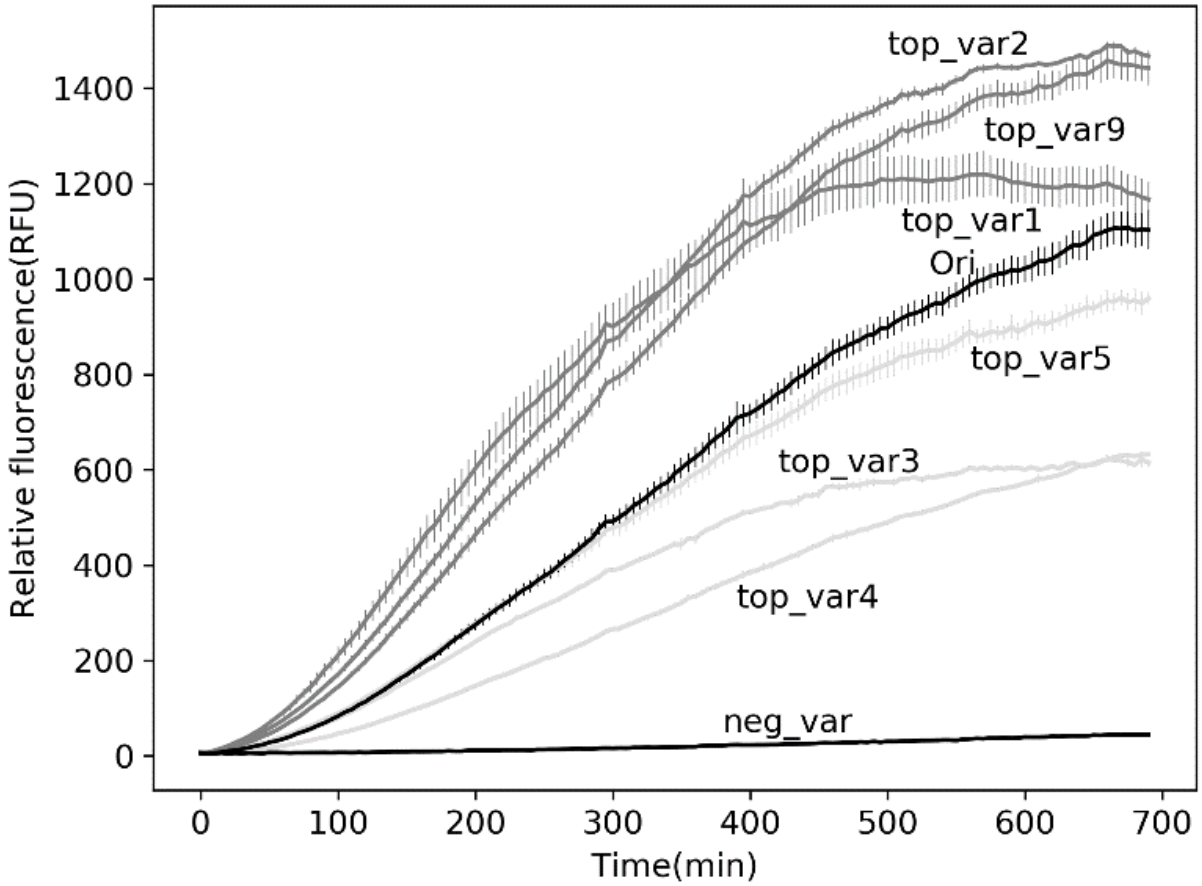


Fig. 4. The design process of GFP11 variants. (A) The strand to design (GFP11) is highlighted in red and query sequences to consider are shown on the right. Dotted lines represent hydrogen bonds. The residues pointing inward, which are not subject to mutation, are colored in gray. (B) The B-SIDER score of the original GFP11 (▼) is compared against a 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 which never appeared in 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.



469 **Fig. 5. Complementation assay with the designed GFP11 variants.** Among the six successfully
470 expressed variants, three variants exhibited stronger fluorescence emissions than the original
471 peptide identified by directed mutagenesis.



472

473

474

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
Alpha B crystalline	3sgo, 4sgp
β-2 microglobulin	3loz, 4e0k, 4e0l
Immunoglobulin	6dj0, 6dix
SOD1	5wor

475

Table 2. GFP 11 variants

476

477

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.