Supporting Information

Zheng et al. 10.1073/pnas.1222130110

SI Text

Simulation Model. The associative-memory, water-mediated, structure and energy model (AWSEM) is an off-lattice coarse-grained protein simulation model. The Hamiltonian is summarized in Eq. **S1**. It consists of a backbone term, V_{backbone}, which restricts the chain to polypeptide-like conformations and is mostly sequence independent, and other terms that depend on the identities of the interacting residues. The burial term, V_{burial}, attempts to sort each residue into its preferred burial environment—exposed, partially buried, or completely buried. The contact term, V_{contact}, consists of a direct contact interaction and a water- or protein-mediated interaction. This terms switches smoothly between the water- and the protein-mediated interaction weights depending on the instantaneous local density of the interacting residues. For the purposes of this study, the hydrogen-bonding term, V_{HB}, consists of two interactions that favor β-hydrogen–bonding geometries. The first one is sequence independent and long range and favors cooperative formation of β-sheets, and the second one is sequence dependent and depends sensitively on the distance and relative orientation of the interacting groups. Finally, the local-in-sequence interactions ($3 \leq |i - j| \leq 9$) are governed by the associative-memory term, V_{AM}. In this study, the associative-memory term is determined by a "single memory," i.e., is a sum of pairwise Gaussian interactions between C_α and C_β atoms that are nearby in sequence with minima at the distances in the experimentally determined native structure. AWSEM is described in detail in the supporting information of a recent paper by Davtyan et al. (1):

$$V_{\text{total}} = V_{\text{backbone}} + V_{\text{contact}} + V_{\text{burial}} + V_{\text{HB}} + V_{\text{AM}}. \qquad \textbf{[S1]}$$

Simulation Protocol. All simulations were performed in the canonical ensemble, using the Nose–Hoover thermostat as implemented in the LAMMPS molecular dynamics package (2). Annealing simulations were started from totally extended structures at a temperature of 450 K and the temperature was cooled over 8,000,000 steps to 350 K with a step size of 5 fs. The energy distributions in Fig. 5 of the main text were generated by taking structures that were found at the end of the simulated annealing simulations and simulating them at 250 K for 200,000 steps.

Frustratometer. The frustration analyses in Fig. 2 of the main text were done using the Frustratometer. The Frustratometer is a tool for localizing frustration in natively folded proteins. It was first described in ref. 3 and is now available as a web server (4). The frustration level of each native interaction is determined by comparing the native interaction energy to the mean of a set of "decoy" interaction energies that would likely be found in compact misfolded states. This "local energy gap" is normalized by the SD of the distribution of decoy energies and the resulting dimensionless quantity is called the frustration index. If the frustration index for a particular interaction is found to be in the minimally frustrated range, it is visualized with a green line. If an interaction has a frustration index in the highly frustrated range, then it is shown in red. Neutrally frustrated interactions are omitted for clarity.

AWSEM-Amylometer. The AWSEM-Amylometer attempts to quantify the propensity of short sequence fragments to form amyloid-like structures. The propensity of a particular six-residue sequence is estimated by threading the sequence onto a "steric zipper" structure [Protein Data Bank (PDB) ID 1YJO] (5) and calculating the AWSEM energy in Eq. **S1**. Structure 1YJO is a crystal structure of a peptide from the sequence of Sup35, a yeast prion protein that is known to form amyloid-like fibrils. The symmetry information in the PDB file and the symexp function in the PyMOL software package (6) were used to create a structure with two sets of three parallel β-strands stacked face-to-face and antiparallel with respect to each other. By threading the entire sequence of a protein, six residues at a time, onto 1YJO and calculating the AWSEM energy each time we obtained an estimate for the relative amyloid propensity of each six-residue fragment in the protein of interest. Similar methods (7) have been used with other energy functions for the special case of all six sequences being identical (as is the case in the original structure). We ran the AWSEM-Amylometer in two distinct modes for this study. In the first mode, self-recognition energies were calculated by threading six identical sequences onto the the structure of 1YJO, similar to what had been done previously (7). In the second mode, a heterogeneous but symmetric structure was constructed by threading two different six-residue sequences such that one β-sheet contains two instances of sequence A separated by one sequence of B and the other contains two of B separated by one of A. An energy was calculated for every possible pair of hexamers A and B of a given protein. In addition, the energy of A and B′ was also calculated, where B′ is the reverse of B, e.g., B = NNQQNY and B′ = YNQQNN. We took this last step to efficiently approximate the energy of antiparallel association of sequences. This is necessary because many β-sheets in the native states of proteins contain antiparallel pairs of β-strands, but this type of arrangement is not directly represented in 1YJO. The second mode allows us to directly compare native-like, nonnative, and self-recognition hexamer interaction energies, giving a more complete picture of the interactions possible during the folding/misfolding process of protein with β-sheets.

A database of hexamer peptides that have been experimentally determined to be either fibril forming or not fibril forming was previously reported (8). To determine an approximate threshold energy, below which a hexamer is likely to be amyloidogenic, we first calculated the self-recognition energies of all 158 experimentally studied peptides, using the method described above. The histograms of energies of the fibril-forming (red) and non-fibril–forming (green) peptides are shown in Fig. S3. We then selected as the threshold energy the value that maximized the number of correctly categorized peptides, i.e., the number of non-fibril–forming peptides above the threshold plus the number of fibril-forming peptides below the threshold. This procedure yielded a threshold value of −262.

1. Davtyan A, et al. (2012) AWSEM-MD: Protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *J Phys Chem B* 116(29):8494–8503.
2. Plimpton S, et al. (1995) Fast parallel algorithms for short-range molecular dynamics. *J Comput Phys* 117:1–19.
3. Ferreiro DU, Hegler JA, Komives EA, Wolynes PG (2007) Localizing frustration in native proteins and protein assemblies. *Proc Natl Acad Sci USA* 104(50):19819–19824.
4. Jenik M, et al. (2012) Protein frustratometer: A tool to localize energetic frustration in protein molecules. *Nucleic Acids Res* 40(Web Server issue):W348–W351.
5. Sawaya MR, et al. (2007) Atomic structures of amyloid cross-beta spines reveal varied steric zippers. *Nature* 447(7143):453–457.
6. Schrödinger L (2010) *The PyMOL Molecular Graphics System, Version 1.3r1* (Schrödinger, Portland, OR).
7. Zhang Z, Chen H, Lai L (2007) Identification of amyloid fibril-forming segments based on structure and residue-based statistical potential. *Bioinformatics* 23(17):2218–2225.
8. Thompson MJ, et al. (2006) The 3D profile method for identifying fibril-forming segments of proteins. *Proc Natl Acad Sci USA* 103(11):4074–4078.

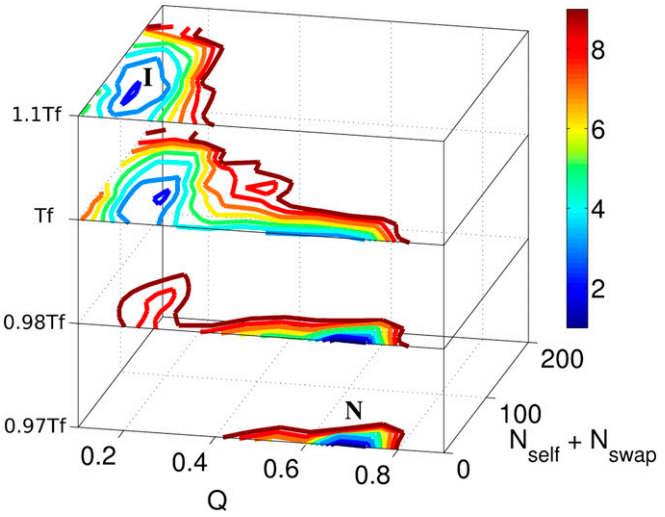

Fig. S1. Free energy curves for I27-I27 for four different temperatures. *N* is the native state, and *I* is the misfolded state stabilized by self-recognition contacts and swapped contacts. As temperature increases, the equilibrium population shifts from the native state to the misfolded state. This indicates that the misfolded ensemble is energetically less favored than the native state, but entropically more favored. *Tf* is the folding temperature.

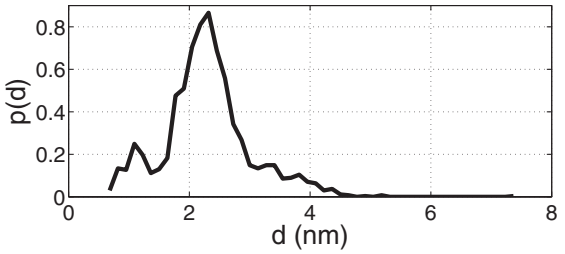

Fig. S2. From simulated annealing simulation of I27-I27 fused dimer, we collected the misfolded structures ($Q^i_{\text{domain}} < 0.4$) near the end of the simulations and calculated the C_α-C_α distance distribution between residues E3 and N83, the two residues where FRET labels were attached in Borgia et al.'s work (1). The curve peaks at 2.3 nm.

1. Borgia MB, et al. (2011) Single-molecule fluorescence reveals sequence-specific misfolding in multidomain proteins. *Nature* 474(7353):662–665.

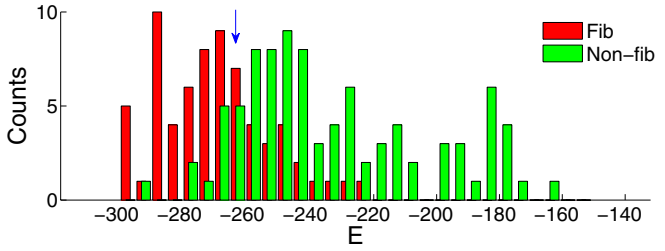

Fig. S3. The energetic threshold for amyloidogenicity for the AWSEM-amylometer was determined by calculations over a dataset of 158 experimentally studied hexamer peptides (8), as explained in the *AWSEM-Amylometer* section. Sixty-seven of the peptides form fibrils and 91 of them do not. The blue arrow indicates the threshold. With this threshold, there are 73% and 89% true positives in fibril-forming peptides and non-fibril–forming peptides, respectively.

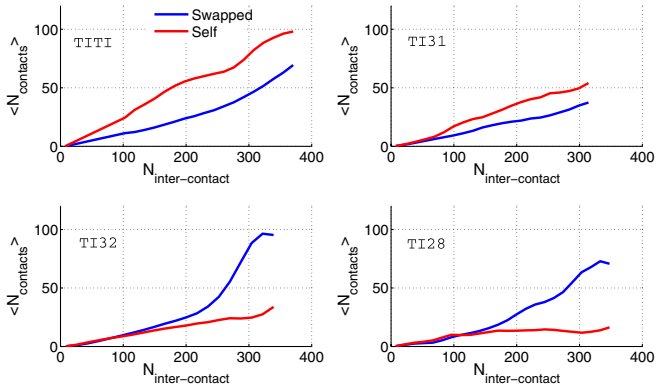

Fig. S4. Competing roles of the swapped contacts and self-recognition contacts in the fused dimers I27-I27 (TITI), I27-I31 (TI31), I27-I32 (TI32), and I27-I28 (TI28). Forty annealing simulations were carried out for all of the fused dimer constructs. $\langle N_{\text{contacts}} \rangle$ is the number of contacts averaged over all 40 simulations. $N_{\text{intercontact}}$ is the number of interdomain contacts. The interdomain contacts include the swapped contacts, self-recognition contacts, and other types of interdomain contacts. As the sequence identity between the two domains decreases, the number of self-recognition contacts decreases monotonically. The number of domain-swapped contacts, however, changes in a more complicated way. On one hand, domain-swapped contacts are forming more often due to the weakening of the self-recognition interactions. On the other hand, lowering the sequence identity decreases the total number of possible domain-swapped contacts.