BIOINFORMATICS

Vol. 00 no. 00 2011
Pages 1–6

A method for probing the mutational landscape of amyloid structure

Supplementary information

1 STRUCTURE PREDICTION ANALYSIS DETAILS

For each protein and protein mutant mentioned, predicted ensemble results are derived from a stochastic sampling of whole structures out of the Boltzmann statistical mechanical ensemble. A summary for each protein tested is given below. From this sampling, conformational populations are clustered by partitioning around mediods (a method similar to k-means), with a fixed number of clusters. To determine how many clusters to fix, we iteratively run the clustering algorithm with an increasing number of clusters until no new major conformational populations are qualitatively differentiated. Since we represent protein structure at the granularity of β-sheet residue/residue hydrogen bond pairs, the distance metric used within the clustering algorithm evaluates two structures' separation according to a sum of (1) the per-residue secondary structure assignment overlap (1-D Hamming distance), (2) the intersection of hydrogen-bonding-pair assignments (an F-measure permitting 'one-off' (i,j)↔(i,j+1), etc. matches), (3) the F-measure statistic of the overlap of full β-strands (where a match occurs even if there is a shift between predicted structures, e.g. a β-strand at positions 60-67 can match another at positions 62-68), and (4) a similar F-measure statistic for the overlap of coil regions. By combining these four criteria, we separate more realistic structural populations than is possible by any of the metrics individually. Finally, a stochastic contact map is calculated from these samples, identifying individual residue-pairs which are highly likely to form β-structure within the ensemble. This can be presented as a heat-map identifying the weighted likelihood $p(i, j)$ that any residue position i will be in contact with position j out of all ensemble structures (Supp. Figure 9). Although an exact calculation of residue-pair likelihoods is supported by our algorithm, a reconstruction of these likelihoods from samples matches this exact value well, and is much less resource-intensive. Optional heuristics can be applied to speed computation by limiting the ensemble via energetic thresholding (Waldispühl *et al.*, 2009)

Amyloid Beta (Aβ): We present the structural ensemble of A$\beta_{1\text{-}42}$ using schema $\mathcal{P}$. We note that both schemas $\mathcal{P}$ and $\mathcal{S}$ (but not schema $\mathcal{A}$) are capable of predicting the exact published single-rung β-solenoid structure of A$\beta_{1\text{-}42}$ due to an intersection in the conformational space defined by each model. However, schema $\mathcal{P}$ defines a much larger space of possible structures, and therefore was chosen to highlight the discriminative power of our scoring function.

To predict an ensemble of A$\beta_{1\text{-}42}$ structures with β-strands of length 6 to 12 using schema $\mathcal{P}$, we first calculate seven sets of sampled structures, fixing a different β-strand length to each, and determine cluster populations across all of those structures combined. This is necessary to ensure a well-distributed coverage of β-strand lengths with fewer samples — predictions that allow the length to vary between 6 and 12 within a single execution explore variations in kink location more often than variations in length. This is due to the two-fold dimensionality increase caused by kinks, and the fact that changes in kink location (and the resulting change in β-strand residue orientations) often can induce a smaller energy difference than changes in β-strand length. From inspection of this data we find the most energetically favorable cluster predominantly contain β-strands of length 9 to 10. Predicting an ensemble of structures based on this length results in two major structural clusters (Supp. Figure 2a), with the largest cluster's mediod structure containing β-strands at positions 17-26 and 32-39, and the smaller at positions 12-24 and 30-40. Although the larger cluster fills 55% of the ensemble and the small 39%, we note that the close similarity between these two clusters of structures may introduce error in calculating a specific percentage value.

The assignment of β-strand at positions 17-26 in the larger cluster and 12-24 in the smaller cluster also agrees with recent experimental studies showing that Aβ fibrils formed under quiescence and agitation differ in their assignment of position 15 to β-strand(Petkova *et al.*, 2005) (agitated fibrils exhibit the Q15 β-strand chemical shift). However, unlike these agitated fibrils, AmyloidMutants predictions in the larger cluster do not show β-strand in positions 10-14. Further, experimental study has shown that brain-seeded fibrils exhibit spatial proximity between residues F19/I31, whereas unseeded *in vitro* fibrils do not(Paravastu *et al.*, 2009). AmyloidMutants also exhibits such a divergence in the predicted ensemble. Predicted β-strands place F19 on the interior (buried) side of the fibril in 89% of the ensemble, while a smaller, but still significant 19% of the ensemble contains a buried I31 — the combined case where both are buried makes up 16% of the ensemble. The predicted assignment of both F19/I31 as buried would allow the kind of spatial proximity observed experimentally.

For illustrative purposes we include predictions of A$\beta_{1\text{-}42}$ using schemas $\mathcal{A}$ and $\mathcal{S}$ (Supp. Figure 2b and Supp. Figure 2c). In these cases similar β-sheet interaction regions present themselves, although the specific β-strand residue/residue pairing can be quite different. This highlights the difficulty in directly comparing schemas against one another.

HET-s: Similar to the procedure used for A$\beta_{1\text{-}42}$, to predict an ensemble of HET-s structures with β-strands of length 6 to 12 using schema $\mathcal{P}$, we first sample and cluster conformations with fixed β-strand length 6, 7, 8, 9, 10, 11, and 12. Inspection reveals that the most energetically favorable cluster of structures predominantly contains β-strands of length 10 to 11. Predicting an ensemble of structures using this length parameter results in two clear populations — the larger grossly corresponding with a

"two-rung-per-chain β-solenoid/β-helix" and the smaller a "one-rung-per-chain" β-solenoid/β-helix" (Supp. Figure 3a). This strong predictive bias toward only two possible structures may relate to the known homogeneity of HET-s fibrils and lack of observed strains(Coustou *et al.*, 1997). Similar predictions were performed for FgHET-s using β-strand lengths of 10 to 11 (Supp. Figure 4).

Again, for illustrative purposes we include predictions of HET-s using schemas $\mathcal{A}$ and $\mathcal{S}$ (Supp. Figure 3b and Supp. Figure 3c). Here we see marked shifts in β-sheet interaction regions between schemas, along with changes in population distribution and β-strand residue/residue pairing.

Amylin: The Boltzmann ensemble for Amylin was computed using schema $\mathcal{S}$ because of this schema's ability to incorporate both 2-sheet β-helical structures as well as 3-sheet serpentine structures within the same conformational space. This allows a comparison between the energetic favorability of 2-sheet versus 3-sheet structures. Supp. Figure 5 presents the predicted ensemble when allowing β-strand length to vary between 6 and 12 residues long (since schema $\mathcal{S}$ does not include kinks). Since residues 1-7 have been shown non-critical to fibril formation due to a disulfide bond between Cys2 and Cys7, we explicitly fix positions 2 and 7 as non-β-sheet-forming, effective throughout all computed structures within the ensemble (such point-wise constraints can be similarly be applied to schemas $\mathcal{P}$ and $\mathcal{A}$). Clustered into three populations, 80% of the ensemble consists of 2-sheet β-solenoid structures with the remaining minority containing 3 sheets. As mentioned, the largest, and mostly homogeneous population contains a β-strand region which overlaps Phe23, a residue shown to interact across peptide chains in a manner unaccounted for by our model(Luca *et al.*, 2007). The second population exhibits much more heterogeneous structure, however, a clear alternative set of conformations exist that avoid β-sheet at Phe23 (as indicated by its mediod).

α-synuclein: Ensemble predictions of α-synuclein were performed using schema $\mathcal{S}$ since this schema permits many β-sheets of differing lengths to pack together without the need for intra-peptide hydrogen bonding interactions. Structures were sampled, allowing β-strands to range in length from 6 to 12, and were clustered into two populations, shown in Supp. Figure 6. Although β-structure regions predicted within the fibril forming region of 30-110 show excellent agreement with experimental observations, two false-positive β-strand structures are apparent around positions 5-10 and 15-20. The likely reason for the prediction of amphipathic β-strands in this region is because this disordered N-terminal is believed to favor a lipid-binding amphipathic α-helix structure(Volles *et al.*, 2007).

Tau (τ): Ensemble predictions of the 441 residue long Tau were performed using schema $\mathcal{S}$ for similar reasons as for α-synuclein. Structures were sampled and clustered into two populations, shown in Supp. Figure 7a (again, permitting β-strands to range from 6 to 12). For demonstrative purposes the number of clusters was fixed to 2, although note that both clusters appear quite similar, suggesting only small β-strand registration variations across the ensemble, and an especially strong consensus on the large regions do not form fibril. Despite the high accuracy in predicting experimentally observed β-strand regions, β-sheet structure is incorrectly predicted around positions 121-128 and 408-430, which overlaps with observed α-helices similar to α-synuclein.

Supp. Figure 7b gives clustered ensemble predictions when the energy model is artificially tuned to bias against β-sheet formation (implemented as a simple scaling factor). Such a calculation can identify peptide regions which have the highest likelihood of forming β-structure. In this case, the mediod of the largest cluster predicts strands at positions 274-279 and 305-310, in agreement with experimental evidence that these regions are crucial in initiating fibril assembly(von Bergan *et al.*, 2000).

2 Aβ / HET-S MUTANT ANALYSIS DETAILS

Provided are more detailed predictions of the conformational shifts identified in the Aβ Iowa mutant and yeast-toxic mutants of HET-s. Similar conformational sampling and clustering was employed as for structural prediction validation in Section 1.

Aβ Iowa mutant: The Boltzmann partition function for both A$\beta_{1\text{-}40}$ and A$\beta_{1\text{-}40/D23N}$ was computed using schema $\mathcal{A}$, and restricting β-strand length between 9 and 10 residues long, conformations were sampled from the ensemble and clustered into two populations. Supp. Figure 10a details the predicted ensemble of A$\beta_{1\text{-}40}$, while Supp. Figure 10b shows that of A$\beta_{1\text{-}40/D23N}$. The composition of these clusters in terms of β-strand registration (e.g. 17+*k*↔21-*k*) is then shown in Table 1.

HET-s yeast-toxic mutants: To illustrate the impact of mutations on the predicted HET-s structure ensemble, conformations were sampled from the predicted Boltzmann ensemble of all five mutants m4, m8, m3, m9, and m11, using both schemas $\mathcal{P}$ and $\mathcal{A}$. Supp. Figure 11 indicates clustered structural populations for both the WT and mutant ensembles under both schema $\mathcal{P}$ (Supp. Figure 11a,b,c,d,e,f) and schema $\mathcal{A}$ (Supp. Figure 11g,h,i,j,k,l). Both calculations assume a β-strand length 10 to 11. Inspecting the results, schema $\mathcal{P}$ appears to permit two major structural populations across all mutants, a "two-rung-per-chain" β-solenoid (with β-structure approximately at positions 7-17, 19-29, 43-53, and 55-65) or a "one-rung-per-chain" β-solenoid (with approximate β-structure positions 21-30 and 41-51). Schema $\mathcal{A}$ permits three major structural populations across mutants, two types of "two-rung-per-chain" β-solenoids, distinguished by the location of their β-strands (approximate β-structure positions of the first type: 2-12, 20-30, 38-38, and 54-64, and approximate positions of the second type: 14-24, 29-39, 43-53, and 58-68), and a "one-rung-per-chain" β-solenoid structure (with approximate β-structure positions 22-32, 51-60). Note, mutant m8 only exhibits one of the two types of "two-rung-per-chain" β-solenoid, seen by the overlap of β-strand positions in the clustered populations. When using schema $\mathcal{P}$, some mutants show dramatic change in the ensemble population, such as m4's increase in propensity to form a "two-rung-per-chain" β-solenoid over a "one-rung-per-chain" β-solenoid, or m8's change from a more homogeneous WT structural population to a more heterogeneous population with many different β-strand/β-strand registrations. Oppositely, schema $\mathcal{A}$ indicates a more heterogeneous population given the WT sequence, and a more ordered population for m8. Note that predictions are conditioned by their schema, and

therefore an interpretation of these predicted results depends on an analysis of the ensemble changes using both schemas.

3 SCHEMA ENSEMBLE CHARACTERISTICS

Mutational occupancy: AmyloidMutants ensembles can describe both structural variation and sequence variation under the assumption of an amyloid fibril fold. Since all sequences result in an amyloid fibril conformation, one means for determining whether a particular mutation makes a peptide more or less amyloidogenic is to quantify each sequence's energetic contribution to the ensemble as a whole. In other words, if only two sequences are permitted for a prediction, the ensemble will contain 50% of the states with one sequence and 50% of the states with the other. However, the energetic weight of the structures resulting from these sequences will vary. In this case, we would assert that the sequence with the larger energetic weight is more amyloidogenic since it forms better energy structures. For instance, when comparing predictions of a WT and mutant sequence, if the mutant sequence occupies 90% of the energetic weight of the ensemble, then the mutant is suggested as a better amyloid forming sequence. Note, however, such comparisons are only valid within a single prediction using one schema, where all possible states are accounted.

Null-hypothesis contact maps: Schemas define a particular set of millions of possible structures, and therefor the likelihood that two residues *i* and *j* within this set form a β-strand hydrogen bond is not uniform across all (*i*,*j*) pairs. To remove this bias from predicted contact maps, we subtract from any predicted $p_{i,j}$ the probability of that (*i*,*j*) pair occurring assuming a constant energy for all interactions within a schema (a "null hypothesis contact map"). This is most easily accomplished by fixing the Boltzmann constant $RT = \infty$ in our energy model. Supp. Figure 13 visualizes this. Reported predictions exhibited a low correlation between predicted null hypothesis contact maps.

4 HET-S AND HET-S/4N→Q MUTAGENESIS

Sequences encoding HET-s WT and HET-s/4N→Q proteins were subcloned into pRH1(Alberti *et al.*, 2009) to allow their expression in bacteria as 7xHis fusions. The proteins were expressed in *E. coli* strain BL21-AI and purified under denaturing conditions, as described(Alberti *et al.*, 2009). Methanol-precipitated proteins were resuspended in 6 M GdnHCl, incubated for 5 min at 95°C, and then filtered through a YM-100 Microcon filter immediately prior to use. Proteins were diluted to 20 μM (corresponding to approximately 60 mM GdnHCl in the assembly reactions) in assembly buffer (5 mM K2HPO4, pH 6.6; 150 mM NaCl; 5 mM EDTA; 2 mM TCEP) and allowed to incubate in 1.5 ml non-binding tubes, with 1000 rpm horizontal agitation, at 23°C for 24 hrs. Because HET-s amyloids are difficult to detect using the amyloid-specific dye, thioflavin T, we instead used detergent insolubility as a measure for amyloid formation. Protein aggregates were detected by passaging the post-incubated reactions through a cellulose acetate membrane, followed by washing with 2% Sarkosyl, essentially as described(Alberti *et al.*, 2009). Retained proteins were visualized by Ponceau-S staining.

REFERENCES

Alberti, S., Halfmann, R., King, O., Kapila, A., and Lindquist, S. (2009) A Systematic Survey Identifies Prions and Illuminates Sequence Features of Prionogenic Proteins. *Cell*, 147:146–158.

Coustou, V., Deleu, C., Saupe, S., and Begueret, J. (1997) The protein product of the het-s heterokaryon incompatibility gene of the fungus podospora anserina behaves as a prion analog. *Proc. Natl. Acad. Sci.*, 94(18):9773–9778.

Kim, W. and Hecht, M. H. (2006) Generic hydrophobic residues are sufficient to promote aggregation of the Alzheimer's Aβ42 peptide. *Proc. Nat. Acad. Sci.*, 103(43):15824–15829.

Kim, W. and Hecht, M. H. (2008) Mutations Enhance the Aggregation Propensity of the Alzheimer's Aβ Peptide. *J. Mol. Biol.*, 377:565–574.

Luca, S., Yau, W., Leapman, R., and Tycko, R. (2007) Peptide Conformation and Supramolecular Organization in Amylin Fibrils: Constraints from Solid State NMR. *Biochemistry*, 46(47):13505–13522.

Paravastu, A., Qahwash, I., Leapman, R., Meredith, S., and Tycko, R. (2009) Seeded growth of β-amyloid fibrils from Alzheimer's brain-derived fibrils produces a distinct fibril structure. *Proc. Natl. Acad. Sci.*, 106(18):7443–7448.

Petkova, A. T. *et al.* (2005) Self-Propagating, Molecular-Level Polymorphism in Alzheimer's β-Amyloid Fibrils. *Science*, 307(5707):262–265.

Volles, M.J. and Lansbury, P.T. Jr. (2007) Relationships between the Sequence of α-Synuclein and its Membrane Affinity, Fibrillization Propensity, and Yeast Toxicity. *J. Mol. Biol.*, 366:1510–1522.

von Bergen, M., *et al.* (2000) Assembly of τ protein into Alzheimer paired helical filaments depends on a local sequence motif (^{306}VQIVYK^{311}) forming β structure. *Proc. Natl. Acad. Sci.*, 97(10):5129–5134.

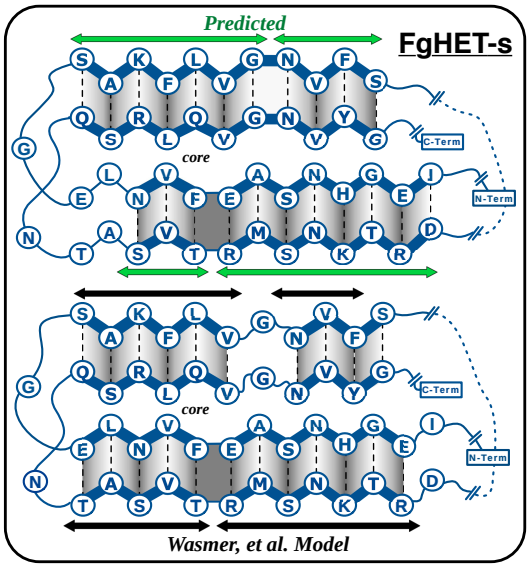
Wasmer, C., *et al.* (2008) Amyloid Fibrils of the HET-s(218-289) Prion Form a β Solenoid with a Triangular Hydrophobic Core. *Science*, 219(5869):1523–6.

Wasmer, C., *et al.* (2010) Structural Similarity between the Prion Domain of HET-s and a Homologue Can Explain Amyloid Cross-Seeding in Spite of Limited Sequence Identity. *J. Mol. Biol.*, 402:311–325.

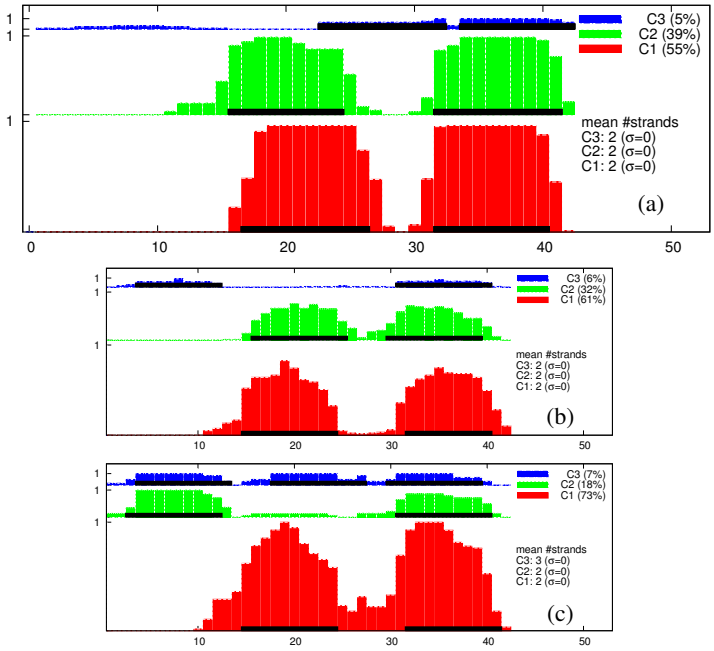
Waldispühl, J., O'Donnell, C. W., Will, S., Devadas, S., Backofen, R., Berger, B. (2009) Simultaneous Alignment and Folding of Protein Sequences. *Proceedings of the 13th Annual International Conference on Research in Computational Molecular Biology (RECOMB).*

Williams, A. D., Shivaprasad, S., Wetzel, R. (2006) Alanine Scanning Mutagenesis of Aβ(1-40) Amyloid Fibril Stability. *J. Mol. Biol.*, 357:1283–1294.

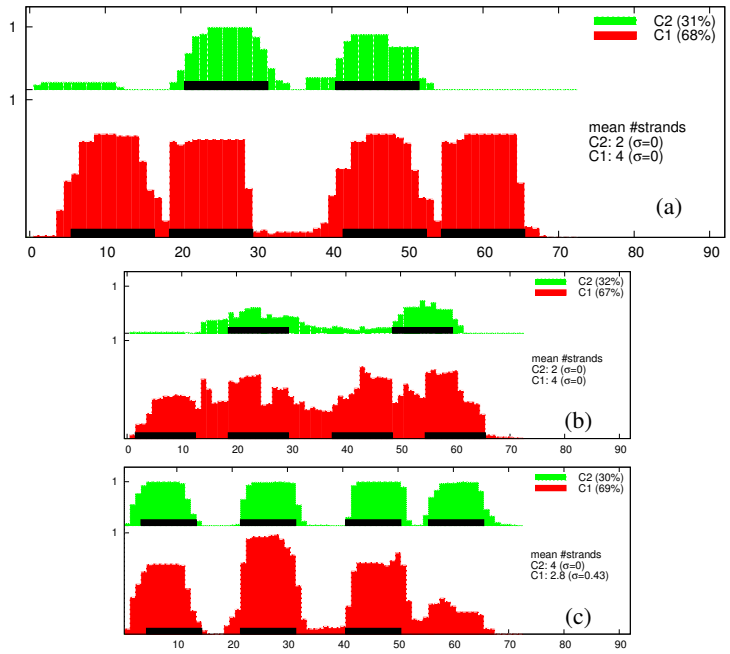
Wurth, C., Guimard, N.K., and Hecht, M.H. (2002) Mutations that Reduce Aggregation of the Alzheimer's Aβ42 Peptide: an Unbiased Search for the Sequence Determinants of Aβ Amyloidogenesis. *J. Mol. Biol.*, 319:1279–1290.



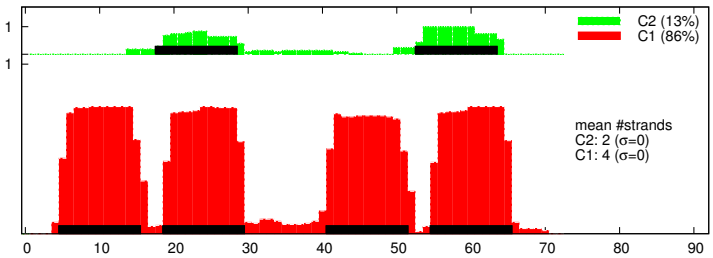
Supp. Fig. 1. AmyloidMutants predictions of FgHET-s closely match a recent partial characterization by solid-state NMR and H/D-exchange(Wasmer *et al.*, 2010). Schema $\mathcal{P}$ was used with strand lengths 10–11. Similar to Figure 3, depicted is is the AmyloidMutants prediction *(top, green arrows)* compared against the putative model *(bottom, black arrows)*, highlighting a near identical match, including residue orientations and and the location of β-strand kinks.



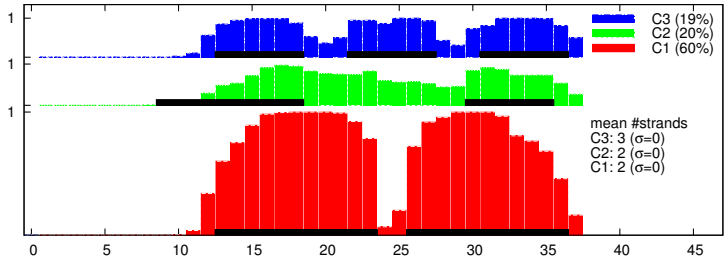
Supp. Fig. 2. **(a)** Schema $\mathcal{P}$ ensemble predictions of A$\beta_{1\text{-}42}$, clustered into three populations. For all such cluster graphs in Supp. Figures 2 through 7, structure cluster populations are separated vertically into colors red, green, and blue, where each cluster's vertical size reflects its energetic weight within the ensemble (given as a percentage in the key). The x-axis indicates residue sequence position while the y-axis indicates the energetically weighted frequency of β-structure at that position within each specific cluster. For example, full bar heights indicate that all structures within the cluster contain a β-strand at that residue position, while a half bar height suggests that half of the energetic weight of that cluster contains structures with a β-strand in that position (i.e. the sum of the energies of the conformations with a β-strand in that position totals half the sum of all conformation energies). The single mediod structure for each cluster has its β-strand regions indicated by a black bar. The average number of β-strands within all structures in each cluster is also indicated. As explained in Section 1, results were attained by sampling conformations with a fixed β-strand length between 9–10. **(b)** Schema $\mathcal{A}$ ensemble predictions of A$\beta_{1\text{-}42}$ using the same parameters. **(c)** Schema $\mathcal{S}$ ensemble predictions of A$\beta_{1\text{-}42}$ using the same parameters.



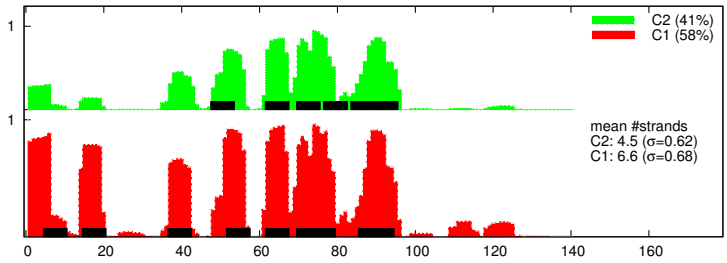
Supp. Fig. 3. **(a)** Schema $\mathcal{P}$ ensemble predictions of HET-s, clustered into two populations, similar to above. As explained in Section 1, results were attained by sampling conformations with a fixed β-strand length between 10–11. **(b)** Schema $\mathcal{A}$ ensemble predictions of HET-s using the same parameters. **(c)** Schema $\mathcal{S}$ ensemble predictions of HET-s using the same parameters.



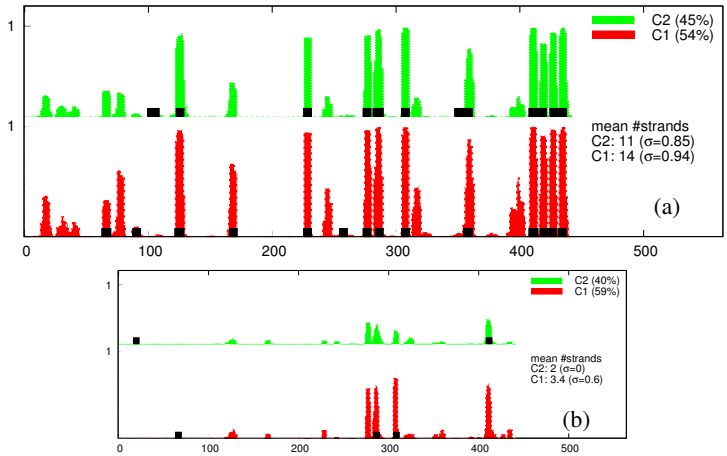
Supp. Fig. 4. Schema $\mathcal{P}$ ensemble predictions of FgHET-s. Results clustered into two populations allowing β-strand length to range between 10–11.



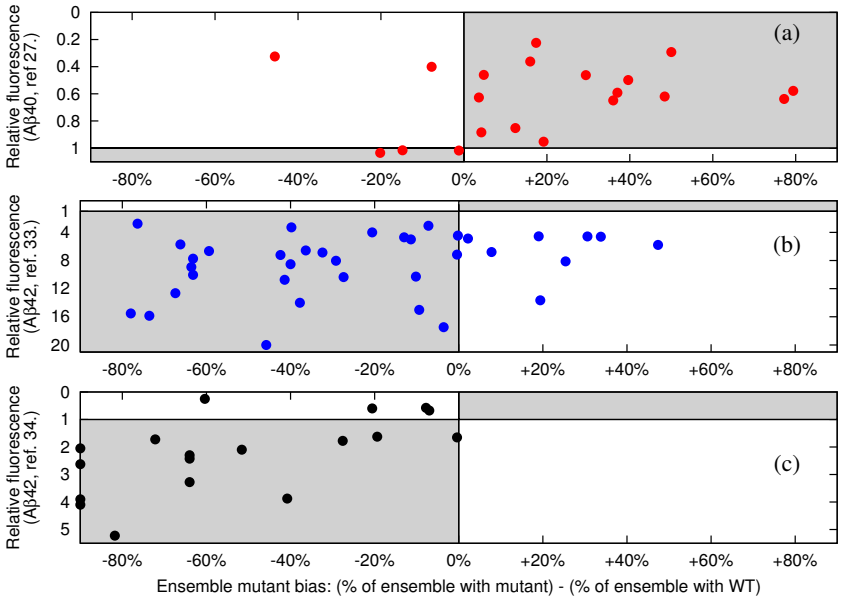
Supp. Fig. 5. Schema $\mathcal{S}$ ensemble predictions of Amylin. Results clustered into three populations allowing β-strand lengths to range between 6–12.



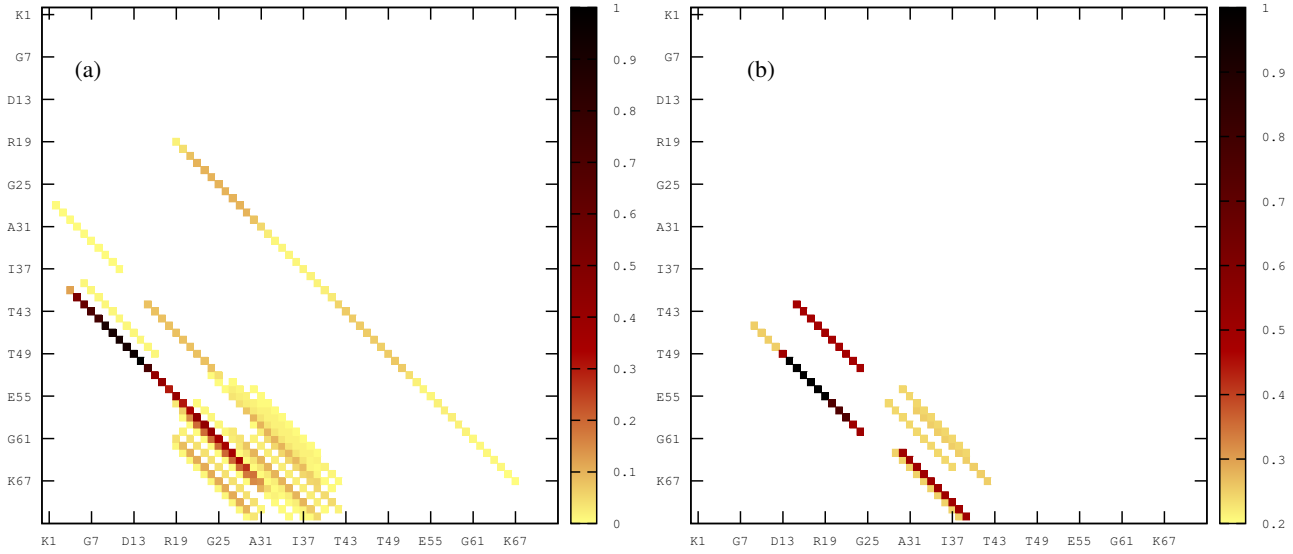
Supp. Fig. 6. Schema $\mathcal{S}$ ensemble predictions of α-synuclein. Results clustered into two populations allowing β-strand lengths to range between 6–12.



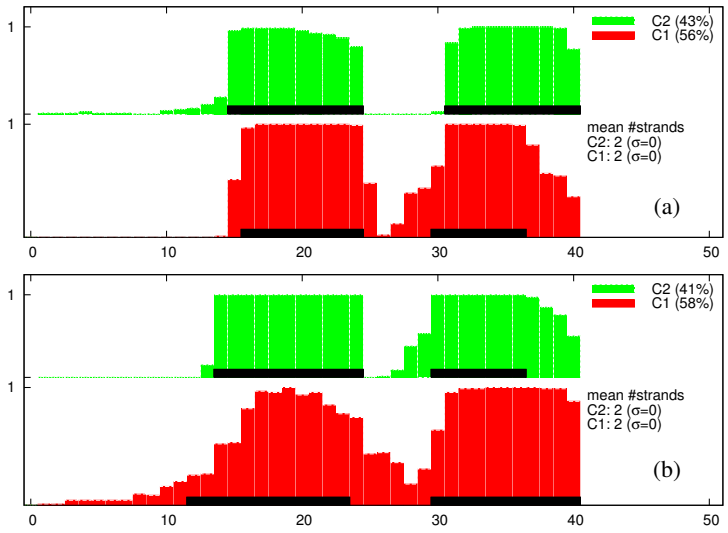
Supp. Fig. 7. Schema $\mathcal{S}$ ensemble predictions of Tau. Results clustered into two populations allowing β-strand lengths to range between 6–12. **(a)** Sampled conformations under the standard energy model. **(b)** Sampled conformations when artificially biasing against β-strand formation within the energy model.



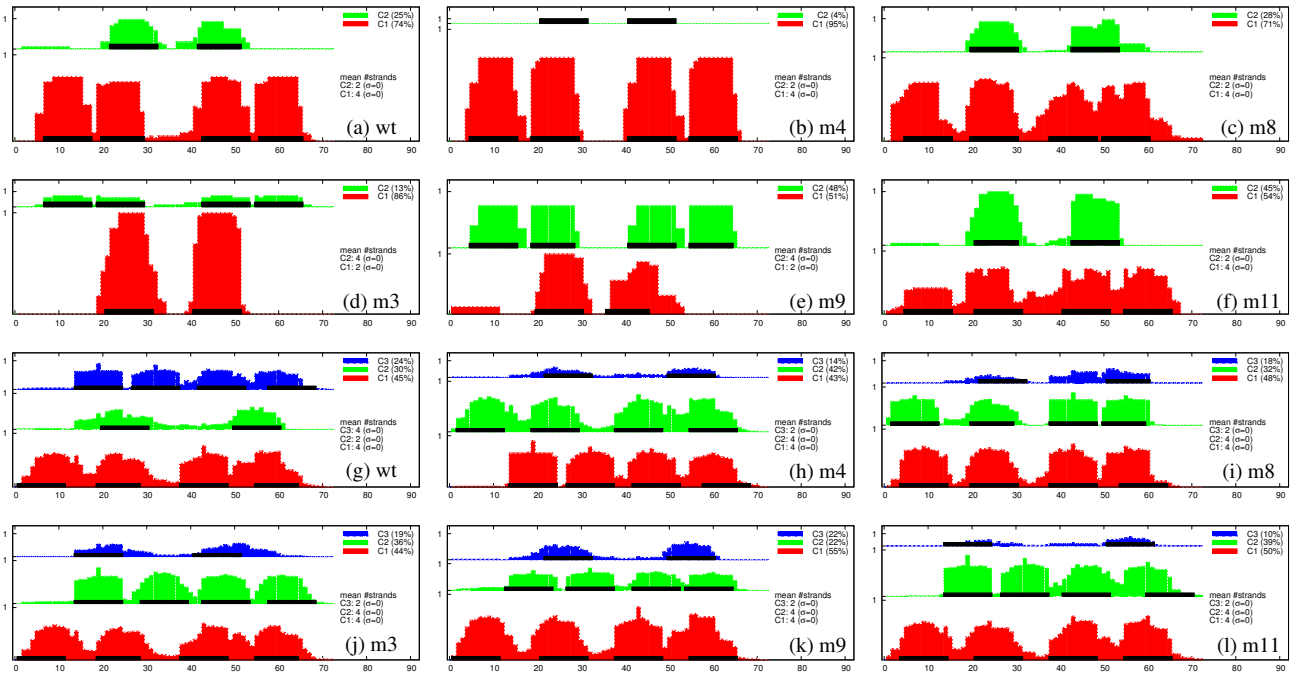
Supp. Fig. 8. AmyloidMutants predictions identify whether a mutant is relatively more or less-amyloidogenic than wild-type. *Postitive x-values* indicate more-amyloidogenic while *negative x-values* indicate less-amyloidogenic than wild-type. Three studies are compared, **(a)**(Kim and Hecht, 2008), **(b)**(Wurth *et al.*, 2002), **(c)**(Kim and Hecht, 2006), that measure mutant vs. wild-type amyloidogenicity by GFP fluorescence — less fluorescence (lower numbers) indicates more amyloid formation. Shaded gray regions indicate when predictions agree with experimental observations.



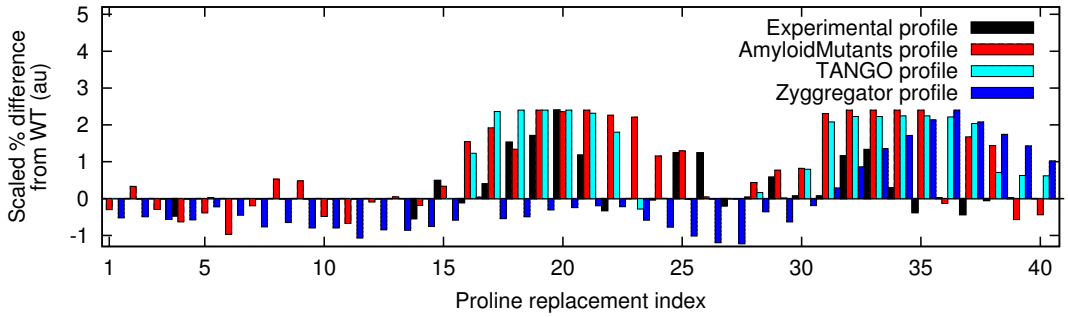
Supp. Fig. 9. Stochastic contact maps highlight conformational differences between AmyloidMutants predictions of WT HET-s and HET-s/4N→Q. Predictions were made using schema $\mathcal{P}$, allowing either WT or HET-s/4N→Q mutations within the ensemble. The stochastic contact map of the WT HET-s cluster is given by **(a)** while the HET-s/4N→Q is given in **(b)**. Sparse contacts in HET-s/4N→Q do not indicate a strong structure prediction, but rather likely result from sampling bias, since HET-s/4N→Q makes up only 4% of the ensemble, or an inability of HET-s/4N→Q to form fibril.



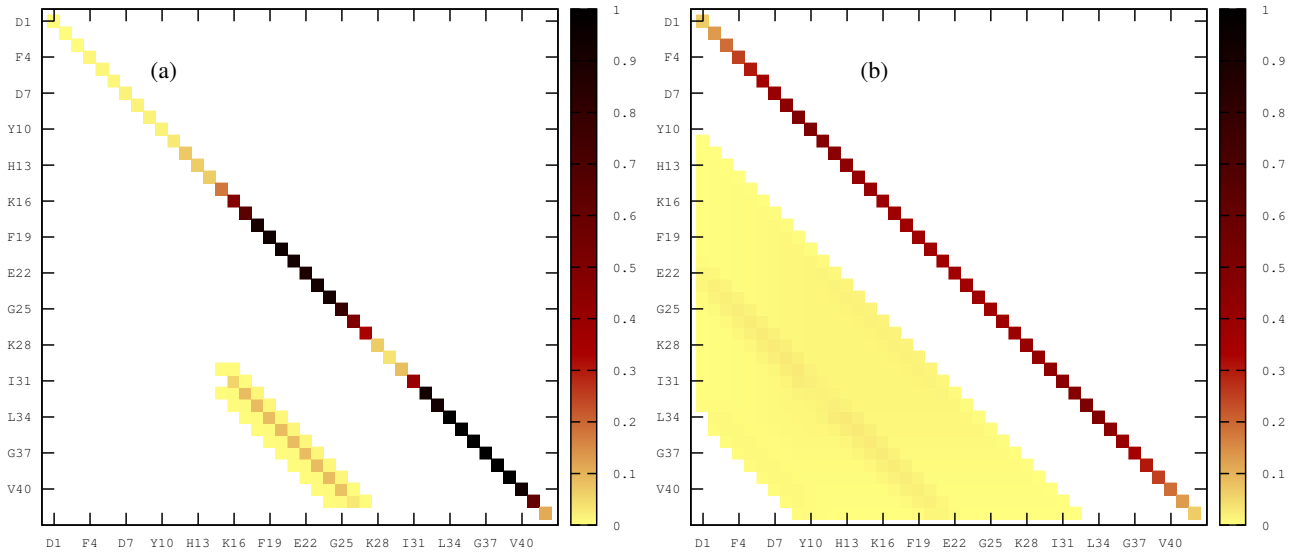
Supp. Fig. 10. Schema $\mathcal{A}$ ensemble predictions of A$\beta_{1\text{-}40}$ and A$\beta_{1\text{-}40/D23N}$. Results clustered into three populations allowing β-strand lengths to range between 9–10. **(a)** A$\beta_{1\text{-}40}$, **(b)** A$\beta_{1\text{-}40/D23N}$.



Supp. Fig. 11. Ensemble predictions of HET-s and HET-s mutants using schemas $\mathcal{A}$ and $\mathcal{P}$. **(a)-(f)** Schema $\mathcal{P}$ predictions clustered into two populations allowing β-strand lengths to range between 10–11. **(g)-(l)** Schema $\mathcal{A}$ predictions clustered into two populations allowing β-strand lengths to range between 10–11. **WT:** (a) and (g). **m4:** (b) and (h). **m8:** (c) and (i). **m3:** (d) and (j). **m9:** (e) and (k). **m11:** (f) and (l).



Supp. Fig. 12. Comparison of AmyloidMutants, TANGO, and Zyggregator to experimental Aβ_{40} scanning mutagenesis data. Experimental "Pro-Ala $\Delta\Delta$G" values indicating the sensitivity of each sequence position to proline replacement (Williams *et al.*, 2006). All three predictors agree with this data around positions 32–33, AmyloidMutants and TANGO also agree with experimental data around positions 18–21, and only AmyloidMutants agrees with experimental data around position 25. Since $\Delta\Delta$G values and each predictor's scores differ greatly in range, results are given in arbitrary units, scaled such that the maximum percent change in aggregation score of any predictor is 2.4 (the maximum experimental $\Delta\Delta$G value). For all bars, positive values suggest that a proline replacement results in a less stable amyloid fibril, while negative values imply the opposite Before scaling, AmyloidMutants values represented the percent difference in ensemble occupancy between WT and mutant sequences, TANGO values represented the inverted percent change in AGG score, and Zyggregator values represented the inverted percent change in Zagg score.



Supp. Fig. 13. AmyloidMutants provides high sequence specificity. To demonstrate this we compare predicted contact maps of HET-s against the null hypothesis contact map (Section 3): **(a)** The predicted contact map of HET-s using schema $\mathcal{P}$ and strand lengths 10–11, **(b)** The null hypothesis contact map of a sequence of the same length using the same schema and parameters. The Pearson's correlation ρ between **(a)** and **(b)** is 0.72 (p=1.6×10^{-145}).