Title

STITCHER: Dynamic assembly of likely amyloid and prion β-structures from secondary structure predictions

Short title

STITCHER: Amyloid β-structure assembly tool

Keywords

structure prediction, β-sheets, energy model, stacking, entropy, dynamic programming

Authors

Allen W. Bryan, Jr. (1) (3) (5)
77 Massachusetts Ave., Cambridge, MA 02139, 617-253-4985, awbryan@mit.edu
Charles W. O'Donnell (3) (5)
77 Massachusetts Ave., Cambridge, MA 02139, 617-253-1450, cwo@mit.edu
Matthew Menke (5)
161 College Ave., Medford, MA 02155, 617-564-4316, mmenke@mit.edu
Lenore J. Cowen (2)
161 College Ave., Medford, MA 02155, 617-627-5134, cowen@cs.tufts.edu
Susan Lindquist (3)
Nine Cambridge Center, Cambridge, MA 02142, 617-258-5184, lindquist_admin@wi.mit.edu
Bonnie Berger (4)(5) *
77 Massachusetts Ave., Cambridge, MA 02139, 617-253-1827, bab@mit.edu

(1) Harvard/MIT Division of Health Science and Technology, Bioinformatics and Integrative Genomics, 77 Massachusetts Avenue, E25-519, Cambridge, MA 02139, USA
(2) Department of Computer Science, Tufts University, 161 College Avenue, Medford, MA 02155, USA
(3) Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02139, USA
(4) Department of Applied Mathematics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, 2-236, Cambridge, MA 02139, USA
(5) MIT Computer Science and Artificial Intelligence Laboratory, The Stata Center, 32 Vassar Street, Cambridge, MA 02139, USA
* Corresponding author

Abstract

The supersecondary structure of amyloids and prions, proteins of intense clinical and biological interest, are difficult to determine by standard experimental or computational means. In addition, significant conformational heterogeneity is known or suspected to exist in many amyloid fibrils. Previous work has demonstrated that probability-based prediction of discrete β-strand pairs can offer insight into these structures. Here, we devise a system of energetic rules that can be used to dynamically assemble these discrete β-strand pairs into complete amyloid β-structures. The STITCHER algorithm progressively 'stitches' strand-pairs into full β-sheets based on a novel free-energy model, incorporating experimentally observed amino-acid side-chain stacking contributions, entropic estimates, and steric restrictions for amyloidal parallel β-sheet construction. A dynamic program computes the top 50 structures and returns both the highest scoring structure and a consensus structure taken by polling this list for common discrete elements. Putative structural heterogeneity can be inferred from sequence regions that compose poorly. Predictions show agreement with experimental models of Alzheimer's amyloid beta peptide and the *Podospora anserina* Het-s prion. Predictions of the HET-s homolog HET-S also reflect experimental observations of poor amyloid formation. We put forward predicted structures for the yeast prion Sup35, suggesting N-terminal structural stability enabled by tyrosine ladders, and C-terminal heterogeneity. Predictions for the Rnq1 prion and alpha-synuclein are also given, identifying a similar mix of homogenous and heterogeneous secondary structure elements. STITCHER provides novel insight into the energetic basis of amyloid structure, provides accurate structure predictions, and can help guide future experimental studies.

Introduction

Amyloid is a highly-ordered cross-β protein aggregate that can be achieved by a very broad set of proteins with widely divergent and unrelated amino acid sequences [1-2]. Given the right conditions, a great many, perhaps most, proteins have the potential to form amyloids. The tendency towards amyloid appears to be due to intrinsic properties of the peptide backbone, a finding of great importance for understanding the evolution of protein folds. A much smaller fraction of proteins, and protein fragments, assemble into amyloid under normal physiological conditions, and these are of great interest in diverse aspects of biology and medicine [3].

Early amyloid research concentrated on amyloids associated with a wide variety of mammalian diseases, from systemic immunoglobulin amyloidoses to neurodegenerative diseases such as Alzheimer's [4]. Initial assumptions that accumulated amyloid always caused cellular and tissue toxicity, as is still believed to take place in peripheral amyloidoses [5], proved to be unfounded upon the discovery of a wider variety of amyloids. Amyloids are now known to play roles in bacterial biofilms [6], the production of melanin [7], the storage of hormones in secretory granules [8], and neuronal learning and memory [9]. A set of self-templating fungal amyloids additionally give rise to epigenetic heritable traits. These bi-stable "prion" proteins can persist as soluble or amyloid species with different functional activities. The self-templating property causes cell-wide persistence of one or the other stable state, a status passed from generation to generation via cytoplasmic transfer of amyloid templates from mother to daughter cells[10-11]. Increasingly, evidence suggests that the formation of amyloids may more commonly be a protective mechanism, which, especially in the case of the neurodegenerative

amyloidoses, acts as to sequester misfolded polypeptides that would otherwise dwell in more toxic, and more highly interactive, oligomeric species. There is, therefore, great interest in deciphering the structures that underlie amyloid states.

While the secondary structure of amyloid is known to be highly β-rich[12-16], experimental structural determination has proven highly difficult, with only extremely short segments crystallized [17-18] and a very few successful solid-state nuclear magnetic resonance (ssNMR) studies [19-23]. Due to the scarcity of direct evidence, the nature of amyloid and prion supersecondary structures and their relation to sequence have been highly contentious topics [17, 24-25]. The parallel β-helices form a fold widely cited as one potential model for amyloid [26-28], while others favor a 'superpleated β-sheet' [29-31]. Complications include the morphological heterogeneity of amyloid structures suggested by EM imagery [27, 32] and the demonstration of prion 'strains' or 'variants' with differing growth and stability phenotypes [33-35]. In the case of the yeast prion protein Sup35, such variants have been demonstrated to maintain specificity through serial passage [34] and have been correlated with differences in conformation [36].

The bi-stable nature of amyloid prions, as well as the observation of heterogeneity and 'strains' in amyloid and prion folding, undermines the canonical viewpoint of 'one protein, one fold' long held by theorists of protein folding. Instead, a murky view arises of a set of stable valleys in a field of conformational configurations, within which variations are permitted around common or similar folding patterns. Enzymologists have long studied the variations in globular protein conformations caused by ligand binding, catalytic activity, presence of ions or cofactors. Amyloids embody a similar but larger set of variations.

Bryan et al. [37] and others have proposed that β-strand *pairs* form the core energetic subunits that make up amyloid structure, and a β-strand predictor was designed around this named BETASCAN. BETASCAN calculates likelihood scores for potential β-strands and strand-pairs based on correlations observed in parallel β-sheets. A key and novel feature of BETASCAN was a maxima-finding algorithm that searched the strands and pairs with the greatest local likelihood for all of the sequence's potential β-structures. While sufficient to predict sequence regions with high potential for amyloidogenic β-structure, BETASCAN did not suggest the most likely overall β-sheet fold. For example, BETASCAN was unable to distinguish between the highly similar amyloidal HET-s allele and non-amyloidal HET-S allele in *Podospora anserine.*

The STITCHER method described in this paper extends prediction of amyloid-like proteins by employing a combination of probabilistic prediction [38] and free-energy [39] methods for protein structural prediction. Since few atomic-detail templates exist from known structures, the algorithm proceeds via a dynamic assembly of β-strands in agreement with the twists and turns necessary to form a β-helix or superpleated-sheet fold. This philosophy of establishing and then manipulating pre-defined structural components has been previously used successfully [40-41]. The score of each successive β-strand addition is determined through a combination of novel free-energy model and BETASCAN-derived probabilities. The free-energy methods account for the enthalpy of created hydrogen bonds and the entropy of linkers, while the probabilities describing the likelihood of β-sheet formation account for the specific sidechain-sidechain interactions that confer structural stability. Of particular importance to our energetic model is the detection of stacking ladders, formed by the sidechain-sidechain stacking and bonding of glutamine, asparagine, tyrosine, and phenylalanine residues [42-44]. To capture the observed structural heterogeneity of amyloids (for example, the "strain" phenomenon), STITCHER calculates a list of the top-scoring M=50 structures instead of just a single optimal. From this set of high-scoring candidates, a consensus structure is derived to represent the commonalities in specific strand-

pairs among these 50 structures. Specifically, portions of the structure are considered more likely to form if they are seen in 80% or more of the top structures.

In our results, we show that the STITCHER method can be used to accurately reconstruct structure, as is given by the example of the well-studied Alzheimer's amyloid beta peptide and the *Podospora anserina* Het-s prion. STITCHER is also shown to be less prone to false positives than the prior BETASCAN program as it distinguishes the amyloid forming HET-s protein from its close, non-amyloidal homolog, HET-S. Novel structural predictions are then analyzed for the prion domain of the yeast protein Sup35 as well as the Rnq1 prion and alpha-synuclein. The STITCHER algorithm may be accessed at http://stitcher.csail.mit.edu.

Materials and Methods

Algorithmic strategy

The greatest problem in protein structure prediction is the reduction of possible conformation patterns from a mind-boggling potential space to the few viable and stable conformations. In the present case, the restriction to parallel, all-β structures massively reduces the conformational space and simplifies the prediction problem. The maxima-finding algorithm of BETASCAN further reduced the space of solutions to only discrete secondary structural elements, identifying locally probable strands and strand-pairs. While this level of detail is useful in some cases, supersecondary and tertiary structural information has been completely lost. The goal of STITCHER is to build upon the successful probability models of BETASCAN to then reconstruct supersecondary structures based on a minimal set of additional energetic features and constraints. While complex hydrophobic and kinetic interactions are known to affect the rate of *in vivo* supersecondary structure formation, STITCHER relies on this minimal set of energy parameters to make the simplest possible model of thermodynamically possible stable structures that could form. Specifically, STITCHER identifies the importance of stacking pairs and entropic penalties for β-strand linkers. Further, β-strands are assembled into full structures using parameters that restrict the length of linkers and the distance between paired strands, parameterized by an interpretation of tertiary β-helix or superpleated-sheet fold. A dynamic program is used to combine ("stitch") BETASCAN strand-pairs using these energetic features and physical constraints, and outputs a list of the top 50 scoring assemblies. Dynamic programs utilize the fact that the calculation of scoring assemblies will re-use many smaller calculations. A piece of pseudocode may demonstrate how STITCHER uses the principle:

```
Calculate_beta_sheet_score (rung_number):
   If lookuptable (beta_sheet_score (rung_number - 1)) exists:
     All_but_last_rung_score =
          lookuptable (beta_sheet_score (rung_number - 1))
   Else
     If rung_number = 0:
        All but last_rung_score =
          Calculate_rung_score (first_rung)
     Else:
        All_but_last_rung_score =
          Calculate_beta_sheet_score (rung number)
     End_if
     Lookuptable (beta_sheet_score (rung_number - 1)) =
          All_but_last_rung_score
```

```
        End if
        Calculate rung_score (last_rung)
        Beta_sheet_score = rung_score + All_but_last_rung_score
        Return Beta_sheet_score
    End Calculate_beta_sheet_score
```

In the pseudocode above, we show a simplified version of a portion of our algorithm calculating a score for a partial structure. First, the algorithm checks if a smaller piece of the structure – namely, all but the last rung – has already been scored. If so, it re-uses that score without the need to repeat a calculation. If it is not scored, it calculates that smaller partial structure's score and stores it. Crucially, for any structure larger than one rung, the algorithm recurses, storing the smaller structures' scores along the way. It then scores the additional rung and stores that score as well. By using a recursive algorithm to make calculations only as needed and storing results for re-use, any partial structure need only be calculated once, saving greatly on calculation time and resources. When no more additional structure is available, the completed structure can then be considered for possible output depending on its score. Once a list of top candidate structures is assembled, we use polling to assess their agreement or disagreement for specific strand-pairs.

Fold constraints and parameters

STITCHER constrains the assembly of strand-pairs to a limited space of amyloid-like parallel β-sheets. Following the conventions of previous authors [26, 37, 45-46] and the evidence from known amyloid models [20-22, 29, 47], we define an arrangement of n possible sheets, discretized by rungs. Each rung contains n strands, each contributing to a sheet. For a structure of m rungs there will be mn strands and $mn - 1$ *linkers* connecting strands to each other. Every rung is connected by strand-pairs of length L, identified by residues *i* and *j* stacked atop one another $\{i, j \geq i, L > 1\}$. Therefore each structure contains $(m-1)n$ strand-pairs in each putative structure. A complete amyloid protofibril is modeled as many copies of such structures.

Bounds on the parameters m and n further reduce the number of choices to be made in selecting plausible structures. The two models of amyloids in the literature may be described by constraints on parameters. The β-helix fold requires $m \geq 1, n \geq 2$ [19, 21, 48]; nearly all observed cases in nature, including solved structures of amyloids, are either $n = 2$ or $n = 3$. A simple model of a superpleated β-sheet is formed by parameters $m = 1, n \geq 2$. In this model every copy of the amyloid protein forms a single rung of *n* β-strands. In the case of $m = 1$, therefore, every strand-pair consists of two identical copies of a strand, and $i = j$ for all strand-pairs. Therefore, two possible sets of parameters were considered for analysis: $n = \{1,2,3\}, m \geq 1$ and $m = 1, n \geq 1, i = j$.

Finally, we restrict the space of potential strand-pair combinations such that $m > 1$ amyloids must loop back to a location near their starting point at the end of each rung. Therefore, the distance from the end of the first strand of any strand-pair to the start of the second must be longer than the strand itself: $(j - i) > 2L$. Likewise, strand-strand linkers can be no shorter than three residues long – the tightest turn observed in crystal and NMR structures.

Free-energy scoring function

Amyloid structures are scored using a formula that includes the BETASCAN scores of their strand-pairs, as well as bonuses reflecting the stabilizing influence of non-backbone hydrogen bonding, and the entropic cost of restricting backbone movement into the loops of strands. The weight of each component in the scoring function was determined by summing estimates of their free-energy contributions to stability. Following the pattern of Zhang *et al.* [49], the Gibbs free energy delta G of the change from unfolded to folded state is

$$\Delta G = \Delta E_c + \Delta E_{el} - T\Delta S_{bb}^{prot}$$

where ΔE_c is the contact enthalpy of placing residues together, ΔE_{el} is the electrostatic energy associated with ionic interactions, and $T\Delta S_{bb}^{prot}$ is the entropy of folding. Amyloids are typically sparse in charged amino acids, and the stronger partial dipoles associated with other amino acids are usually incorporated into hydrogen bonding. In this analysis, the electrostatic interaction is therefore only considered with reference to hydrogen bonds, and ΔE_{el} is set to zero. Contact energies can be further decomposed into energies contributed by backbone-backbone, backbone-sidechain, and sidechain-sidechain interactions:

$$\Delta G = \Delta E_{bb-bb} + \Delta E_{bb-sc} + \Delta E_{sc-sc} - T\Delta S_{bb}^{prot}$$

Role of BETASCAN scores

We model amyloid peptide backbones to contain only β-strands, with a linear arrangement constrained by the hydrogen bonds to other strands, and linkers, which are only constrained by the strands at their beginning and end. There is one hydrogen bond per residue in the length of each strand-pair, and an equivalent entropy loss for the constraint it imposes on backbone flexibility. Therefore we separate the entropy into linker and strand terms, and rearrange to express them with reference to length:

$$\Delta G = (\Delta E_{bb} - T\Delta S_{str.}) + \Delta E_{bb-sc} + \Delta E_{sc-sc} - T\Delta S_{link}$$

$$\Delta G = (\Delta E_{H-bond} - T\Delta S_{bb-res.})L_{str.} + \Delta E_{bb-sc} + \Delta E_{sc-sc} - T\Delta S_{link}$$

Side-chain/backbone interactions are a primary determinant of β-sheet propensities, both through van der Waals interactions [50] and entropy of solvation [51]. These factors explain much of the known relative affinities of amino acids [52], although these propensities must be interpreted in context [53]. The BETASCAN algorithm uses these propensities to estimate relative probabilities of formation of a strand-pair, normalized by length to allow comparison of strands with different lengths. This log-odds estimate of the relative probability of a β-strand conformation is applied to $\Delta G_{\beta-form}$. The energies from the direct hydrogen bonds made by all residues in the strand may be combined with the side-chain effects and estimated by the entire strand score:

$$\Delta G = \sum_{strands} [(\Delta E_{H-bond} + \Delta E_{bb-res} - T\Delta S_{bb-res.})L_{str.}] + \Delta E_{sc-sc} - T\Delta S_{link}$$

$$\Delta G = \sum_{strands} \left[Score\left(\overline{\Delta G_{\beta\, form.}^{per\ res.}}\right) L_{str.}\right] + \Delta E_{sc-sc} - T\Delta S_{link}$$

Side-chain stability bonuses

Side-chain/side-chain energies include hydrogen bonding in the case of asparagine and glutamine stacking [42], pi-bond orbital stacking in the case of tyrosine and phenylalanine [54], and van der Waals interactions between side-chains. The first two contribute bonus stability beyond that calculated by

BETASCAN. The last has been shown to be very small (less than 0.20 kcal/mol [49]) and can be disregarded. We therefore set:

$$\Delta E_{sc-sc} = \sum_{X=\{Q,N,Y,F\}} (n_{XX}) \Delta E_{XX-stack}$$

Entropic penalties

Finally, we consider the entropy of the linkers, defined as free loops of peptide between β-strands. We note, but do not include here, the difficult problem posed by the paradoxical contributions of polyglutamines [55] to the entropy of β-structures such as the huntingtin fibril. The problem of calculating the linker entropy is otherwise a subset of the general problem of polymer condensation entropy [56] and bears remarkable similarity to that of disulfide bond entropy [57]. The entropy may be calculated as

$$\Delta S = -R \ln \left(\frac{3}{\left(2\pi {l_{aa}}^2 L_{link}\right)^{3/2}} \right) v_{ends}$$

where R is the gas constant, l_{aa} the length from α-carbon to α-carbon, 3.8 angstroms, L_{link} is the number of residues in the linker, and v_{ends} is the volume the ends of the linker may occupy. A hydrogen bond is approximately the same length as the distance between sulfide groups in a disulfide bond, namely 4.8 angstroms. The entropy calculation, using these values, may be simplified to

$$T\Delta S = -2.1 \frac{\text{kcal}}{\text{mol} - \text{res}} - \frac{3}{2} R \ln L_{link}$$

for $T \approx 300$ K. Because we are comparing structures known to have linkers, we disregard the constant term and make an estimate yielding the relative entropy of a linker,

$$T\Delta\Delta S = (0.9 \frac{\text{kcal}}{\text{mol} - \text{res}}) \ln L_{link}$$

This formula is used in two different ways: to calculate the entropic penalty of adding β-strands extending a sheet, and to calculate the entropic penalties accrued from combining multiple sheets into the fiber. The difference is in the value of L_{link}. For the former, we assess the entropy of forming a loop from a free polypeptide chain without regard to strand-pairs (as the strand-pairs cannot form until the chain is in proximity to itself). In this case $L_{link}^{rung} = j_1 - i_1$ is the difference between the N-termini of the two strands. For linkers between two separate β-sheets *f* and *g*, the length of the linker is the number of residues between strand-pairs, counting both the upper and lower chains of the pair:

$$L_{link}^{f-g} = i_g - (i_f + L_f) + j_g - (j_f + L_f)$$

The form of the scoring function for STITCHER may now be fully described as

$$\Delta G = \sum_{strands} \left[Score\left(\overline{\Delta G_{\beta\, form.}^{per\ res.}} \right) L_{str.} \right] + \sum_{X=\{Q,N,Y,F\}} (n_{XX}) \Delta E_{XX-stack}$$
$$- \sum_{rungs} \left(0.9 \frac{\text{kcal}}{\text{mol} - \text{res}} \right) \ln L_{link}^{rung} - \sum_{links} \left(0.9 \frac{\text{kcal}}{\text{mol} - \text{res}} \right) \ln L_{link}^{f-g}$$

Energy weights

To calculate this function, we must estimate $\Delta E_{XX-stack}$. Experimental data [49-51, 53] suggests the free energy of β-strand formation per residue to be approximately 1 kcal/mol-res, a combination of the enthalpy of the hydrogen bond and the entropy of solvation as influenced by side-chains. This is a somewhat rough estimate due to context-dependency [53]. For the bonuses and penalties, we assess the contribution of additional hydrogen bonds to the free energy. The free energy of the hydrogen bond is again offset to some degree by solvation, though not as strongly as for the backbone. The rough bonus weights $\Delta E_{QQ-stack} \approx 1\frac{\text{kcal}}{\text{mol-res}}, \Delta E_{NN-stack} \approx 2\frac{\text{kcal}}{\text{mol-res}}, \Delta E_{YY-stack} \approx 1\frac{\text{kcal}}{\text{mol-res}}$ were used for this study. The extra weighting of *NN* over *QQ* is justified in two ways. First, asparagine (*N*) has a shorter distance from backbone to amide than does glutamine *(Q)*. Additionally, experimental data [43] suggests that in at least one prion, replacing all glutamines with asparagines provides better stability of the prion fold as compared to replacing all asparagines with glutamines. These estimated energy weightings may become more accurate as calorimetry of sidechain-sidechain interactions becomes available.

Evaluation of score

The STITCHER algorithm uses a dynamic programming algorithm to evaluate estimated ΔG for the combinations of strand-pairs that can be combined into templates matching the parameters described above. To do so, the calculation of ΔG is subdivided into calculation by rungs. The total free energy change can then be calculated by summing the stability contribution of any rung r containing strands $\{r_1 \dots r_n\}$ and linkers $\{r_{link}^{1-2} \dots r_{link}^{(n-1)-n}\}$, with a linker r_{link}^{0-1} to the previous rung, as

$$\Delta G = \sum_{r=\{1\dots m\}} [\Delta G_r]$$

$$\Delta G_r = \sum_{str=\{r_1\dots r_n\}} \left[Score\left(\overline{\Delta G_{\beta\, form.}^{per\ res.}}\right) L_{str}\right] + \sum_{X=\{Q,N,Y,F\}} (n_{XX}^r)\Delta E_{XX-stack} - \sum_{g=\{1\dots n\}} \left(0.9\frac{\text{kcal}}{\text{mol}-\text{res}}\right) \ln\left(L_{link}^{(g-1)-g}\right)$$

If the stacking bonuses, a directly sequence-dependent calculation, are considered apart from the strand and linker scores, which are only indirectly sequence-dependent, then the rung calculation can be partially separated into strand calculations:

$$\Delta G_r = \sum_{g=\{1\dots n\}} \left[Score\left(\overline{\Delta G_{\beta\, form.}^{per\ res.}}\right) L_{r_g} - \left(0.9\frac{\text{kcal}}{\text{mol}-\text{res}}\right) \ln\left(L_{link}^{(g-1)-g}\right)\right] + \sum_{X=\{Q,N,Y,F\}} (n_{XX}^r)\Delta E_{XX-stack}$$

By calculating the free-energy scores of strand-pairs and linkers as subproblems of rung scoring, and rung scores as subproblems of structure assembly, the dynamic programming method can be used to iteratively calculate the *M* structures with the highest score by tracing back through internally consistent partial structures that do not violate the defined fold constraints.

Evaluating consensus outputs

The composition of the *M* highest structures (with default *M* of 50) is assessed by scanning over all structures for included strand-pairs by the locations (i, j) of their termini. If the number of strand-pairs in the *M* highest structures with N-termini of $(i \pm 2, j \pm 2)$ total more than 80% of M, the location is noted as a consensus structure element, and the ranges of $i, j, L,$ and strand-pair score over the strand-pairs in the $(i \pm 2, j \pm 2)$ region are output. For display purposes, the strand-pairs with matching lower and upper strands are aligned vertically to reconstruct predicted β-sheets. The output of STITCHER includes the set of *M* predicted structures, a diagram of local structure space, and the top scoring consensus structure.

Results and Discussion

Amyloid-beta:

STITCHER was tested on amyloid beta, an amyloid with two experimental NMR models [20, 47], allowing both superpleated sheet $(m = 1, i = j)$ and β helix $(m > 1)$ structures (see Symbols Used for definitions of all variables). The results are shown in Figure 1. In the case of amyloid beta, the highest-scoring structures all incorporate at least one $i = j$ strand-pair (the ten-residue β-strand from isoleucine 31 to valine 40 inclusive). Several of the highest-scoring structures include strands analogous to those in solved NMR structures, including strands beginning at tyrosine 10 (corresponding to [47]) and at leucine 17 (corresponding to [20]). However, the highest-scoring structures include a first strand with one residue shifted from a perfect in-register parallel alignment, in the region between histidine 13 and alanine 21 (sequence HHQKLVFFA). This region is known to exhibit structural variability, especially at differing pH [58].

Het-s/S:

A key goal in amyloid folding studies is to distinguish amyloidogenic from non-amyloidogenic sequences. STITCHER was run on the small-s and big-S variants of the Het-s mating compatibility factor to test its ability to make this classification. The small-s allele of this protein is known to form an amyloidal prion [21-22]. In contrast, the big-S allele does not form an amyloid structure.

The results for Het-s and Het-S are displayed as Figure 2A and Figure 2B, respectively. Immediately evident is the greater "connectivity" of Het-s predictions (i.e. the high number of valid sequential orderings of β-strand pairs). On the other hand, very few of the predicted Het-S strand-pairs are able to form multiple-sheet structures. Furthermore, the single HET-s strand-pair most often seen in the list of 50 top structures (identified by N-terminal residue-pair isoleucine 14 – threonine 49) overlap with the conformation of strands 1b, 2a, 3b, and 4a in the most recently solved NMR structure [22], although off by one in pairing registry.

Sup35p: case study in three species

We used STITCHER to computationally investigate the amyloidogenic impact of minor sequence differences in homologous proteins. Three homologs of the yeast prion Sup35p were chosen, taken from *Candida albicans, Saccharomyces cerevisiae,* and *Yarrowia lipolytica*. The outputs of these predictions are shown in Figure 3.

The top 50 structural predictions for the *C. albicans* sequence suggest only a single possible folding route (Figure 3A), as all top structures contain recurrent β-strand pairs. The consensus structure thus stretches nearly the entire sequence. Further, the probabilistic signal from each structural strand-pair is fairly weak, with most strands consisting of only 3 or fewer residues. It appears that most of the amyloidogenic potential detected in this sequence comes from its long polyglutamine stretches. STITCHER accounts for these via loop-based stacking bonuses that stabilize β-sheet structure. Specifically, tyrosine and phenylalanine ladders serve to align the structure, in agreement with some mutagenic experiments [42].

The top 50 structural predictions for Sup35p *S. cerevisiae* (Figure 3B) and *Y. lipolytica* (Figure 3C) are more heterogeneous. Much of this heterogeneity is due to the recurrent repeats in the Sup35p sequence. For instance, the strand beginning at glycine 44 can favorably be paired with the repeats at glycines 68, 77, and 97. Three pairs in the *S. cerevisiae* structure, identified by their N-terminal residues as asparagine 12 – glutamine 38, tyrosine 29 – tyrosine 49, and glutamine 38 - glutamine 62, are highly recurrent. Interestingly, these pairs correspond to the "head" area identified in previous experimental studies [36], where mutations are known to have large effects on amyloidogenic potential [59]. In contrast, predictions for the region beyond residue 91 shows increased variance. This "tail" region also exhibits experimental variability in different protein "strains" [36]. The *Y. lipolytica* structure (Figure 3C) shows similar traits, but with fewer recurrent β-structures most likely as a result of the irregularity in sequence repeats.

The effects of ladders and residue-repeats in Alpha-synuclein and Rnq1p

To further study the effect of residue sidechain stacking, STITCHER was also used predict the structures of two other important amyloidogenic proteins: human alpha-synuclein and the *S. cerevisiae* prion Rnq1p (Figure 4). Although a putative structural model has been published for alpha-synuclein [60], no structure has been proposed for Rnq1p. However, it has been shown that Rnq1p may facilitate templating in Sup35p fibers [61-62].

The top 50 structure predictions for Rnq1p (Figure 4A) follow a similar pattern as in *C. albicans* Sup35p, identifying many recurrent β-structures. Although nonspecific loop-based sidechain ladders of glutamines contribute the most to the energetic score of these structures, shorter strands including glutamine, phenylalanine, tyrosine, and asparagine ladders also contribute to the identification of specific β-strand pairs. β-structure content is highest in the region of residues 53-175, albeit with large exterior loops exposing a high number of glutamine and asparagine residues. This suggests that polyasparagine and polyglutamine stretches limit highly specific β-strand pairing.

β-helical STITCHER predictions for alpha-synuclein (Figure 4B) output a strong signal for recurrent strand-pairs in the region roughly between residues 20 and 80. The highest-scoring structures are composed of two β-sheets; the first with strands of length 3 and the second with strands of lengths 7 to 10. Interestingly, some strand pairing regions exhibit alternate registries in nearly the same locale and with nearly equal scores. As with amyloid beta, this predicted variability is associated with the existence of repeated residues found near the edge of the strands (for example, valines 15, 16, alanines 17-19; alanines 29, 30; valines 48, 49, glycines 67, 68, valines 70, 71, alanines 88-91, and lysines 96, 97).

Conclusion

STITCHER introduces a novel energetic scoring model for amyloid fibrils and an efficient algorithm for dynamically assembling discrete β-strand pairs into complete amyloid structures. The system of physical constraints used to "stitch" β-strands into a complete structure offers an accurate generalization of successful template-based methods such as BETAWRAP and its successors [44,51,52], which only conformed to rigid templates. In addition, STITCHER takes into account the unavoidable uncertainty in free-energy parameters and the potential heterogeneity of amyloid folds by computing multiple solutions instead a single optimal. Although the highest-scoring fold frequently offers a good solution, the best results are achieved by assembling the top solution's most frequently occurring strand-pairs into consensus structures. It should be noted that the particular fold taken by an amyloid is sensitive to environmental conditions, including pH, concentration of protein, and presence of chaperone proteins. Thus, the structure with the highest STITCHER score may not be that taken by the protein under experimental conditions.

The results for alpha-synuclein, Rnq1p, and to some extent the various Sup35p proteins highlight the role of single-residue repeats, motif repeats, and sidechain stacking ladders in amyloid structure. Single-residue repeats may contribute to structural stability through stacking ladders, especially in the cases of polyglutamine and polyasparagine. However, this stability increase comes at the expense of β-strand pairing specificity, as the importance of aligning any particular pair of residues is reduced. A similar but diminished effect is seen in repeats composed of multiple-residue motifs, such as in *S. cerevisiae* Sup35p. This suggests that the β-strand pairing specificity for repeat-heavy amyloidogenic proteins may be conveyed through two other features of a structure: short intervening linker loops, and the formation of strand-pairs with more rare stacking ladders such as histidine and phenylalanine.

In the cases of known amyloid structures amyloid beta and HET-s, STITCHER is able to predict the core regions of β-structure observed experimentally. Conversely, the predicted results for Het-S agree with the protein's observed non-amyloidogenic nature, despite a high β-propensity sequence that is nearly the same as the amyloidogenic HET-s prion. While a more robust analysis and verification would require additional amyloid structure determination, the STITCHER methodology appears to be a valuable addition to the growing number of amyloid detection algorithms. Further, as new experimental data provides better insight into the nature of β-strand energetics, and new amyloid structures arise, the STITCHER algorithm could be readily extended. While some interpretation and experimental verification is necessary for a complete understanding of amyloid folding, the identification of the range of most likely folds should greatly enhance the further computational and experimental investigation of amyloid and prion proteins.

References

1. Dobson, C., *The structural basis of protein folding and its links with human disease.* Philosophical Transactions: Biological Sciences, 2001. **356**(1406): p. 133-145.
2. Selkoe, D., *Folding proteins in fatal ways.* Nature, 2003. **426**: p. 900-904.
3. Dobson, C., *Protein folding and misfolding.* Nature, 2003. **426**(6968): p. 884-890.
4. Prusiner, S.B., *Prion Biology and Diseases*. 2004, New York: Cold Spring Harbor Laboratory Press.
5. Bucciantini, M., et al., *Inherent cytotoxicity of aggregates implies a common origin for protein misfolding diseases.* Nature, 2002. **416**: p. 507-511.
6. Chapman, M.R., et al., *Role of Escherichia coli curli operons in directing amyloid fiber formation.* Science (New York, N.Y.), 2002. **295**(5556): p. 851-5.
7. Fowler, D.M., et al., *Functional amyloid formation within mammalian tissue.* PLoS biology, 2006. **4**(1): p. e6.
8. Maji, S.K., et al., *Functional Amyloids As Natural Storage of Peptide Hormones in Pituitary Secretory Granules.* Science, 2009. **325**(5938): p. 328-332.
9. Si, K., S. Lindquist, and E.R. Kandel, *A neuronal isoform of the aplysia CPEB has prion-like properties.* Cell, 2003. **115**(7): p. 879-91.
10. Wickner, R.B., et al., *Prions of fungi: inherited structures and biological roles.* Nat Rev Microbiol, 2007. **5(8)**: p. 611-8.
11. Uptain, S.M. and S. Lindquist, *Prions as protein-based genetic elements.* Annual review of microbiology, 2002. **56**: p. 703-41.
12. Sunde, M. and C. Blake, *The structure of amyloid fibrils by electron microscopy and X-ray diffraction.* Advances in protein chemistry, 1997. **50**: p. 123-59.
13. Maddelein, M.L., et al., *Amyloid aggregates of the HET-s prion protein are infectious.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(11): p. 7402-7.
14. Cascio, M., P.A. Glazer, and B.A. Wallace, *The secondary structure of human amyloid deposits as determined by circular dichroism spectroscopy.* Biochemical and biophysical research communications, 1989. **162**(3): p. 1162-6.
15. Soto, C. and E.M. Castaäno, *The conformation of Alzheimer's beta peptide determines the rate of amyloid formation and its resistance to proteolysis.* The Biochemical journal, 1996. **314**: p. 701-7.
16. Kajava, A.V., J.M. Squire, and D.A. Parry, *Beta-structures in fibrous proteins.* Advances in protein chemistry, 2006. **73**: p. 1-15.
17. Nelson, R., et al., *Structure of the cross-beta spine of amyloid-like fibrils.* Nature, 2005. **435**(7043): p. 773-8.
18. Sawaya, M.R., et al., *Atomic structures of amyloid cross-beta spines reveal varied steric zippers.* Nature, 2007. **447**(7143): p. 453-7.

19. Petkova, A.T., et al., *Solid state NMR reveals a pH-dependent antiparallel beta-sheet registry in fibrils formed by a beta-amyloid peptide.* Journal of molecular biology, 2004. **335**(1): p. 247-60.
20. Lèuhrs, T., et al., *3D structure of Alzheimer's amyloid-beta(1-42) fibrils.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(48): p. 17342-7.
21. Ritter, C., et al., *Correlation of structural elements and infectivity of the HET-s prion.* Nature, 2005. **435**(7043): p. 844-8.
22. Wasmer, C., et al., *Amyloid fibrils of the HET-s(218-289) prion form a beta solenoid with a triangular hydrophobic core.* Science, 2008. **319(5869)**: p. 1523-6.
23. Wickner, R., F. Dyda, and R. Tycko, *Amyloid of Rnq1p, the basis of the [PIN+] prion, has a parallel in-register {beta}-sheet structure.* Proceedings of the National Academy of Sciences, 2008. **105**(7): p. 2403.
24. Lansbury Jr, P., *In pursuit of the molecular structure of amyloid plaque: new technology provides unexpected and critical information.* Biochemistry, 1992. **31**(30): p. 6865-6870.
25. Serpell, L., *Alzheimer's amyloid fibrils: structure and assembly.* BBA-Molecular Basis of Disease, 2000. **1502**(1): p. 16-30.
26. Perutz, M., et al., *Amyloid fibers are water-filled nanotubes.* Proceedings of the National Academy of Sciences, 2002. **99**(8): p. 5591.
27. Wetzel, R., *Ideas of Order for Amyloid Fibril Structure.* Structure, 2002. **10**(8): p. 1031-1036.
28. Wille, H., et al., *Structural studies of the scrapie prion protein by electron crystallography.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(6): p. 3563-3568.
29. Kajava, A.V., et al., *A model for Ure2p prion filaments and other amyloids: the parallel superpleated beta-structure.* Proc Natl Acad Sci U S A, 2004. **101**(21): p. 7885-90.
30. Kajava, A.V., U. Aebi, and A.C. Steven, *The parallel superpleated beta-structure as a model for amyloid fibrils of human amylin.* Journal of molecular biology, 2005. **348**(2): p. 247-52.
31. Shewmaker, F., R.B. Wickner, and R. Tycko, *Amyloid of the prion domain of Sup35p has an in-register parallel beta-sheet structure.* Proc Natl Acad Sci U S A, 2006. **103**(52): p. 19754-9.
32. Sipe, J. and A. Cohen, *Review: History of the Amyloid Fibril.* Journal of Structural Biology, 2000. **130**(2-3): p. 88-98.
33. DePace, A.H. and J.S. Weissman, *Origins and kinetic consequences of diversity in Sup35 yeast prion fibers.* Nature structural biology, 2002. **9**(5): p. 389-96.
34. Tanaka, M., et al., *Conformational variations in an infectious protein determine prion strain differences.* Nature, 2004. **428**(6980): p. 323-8.
35. Tessier, P. and S. Lindquist, *Prion recognition elements govern nucleation, strain specificity and species barriers.* Nature, 2007. **447**(7144): p. 556-561.
36. Krishnan, R. and S. Lindquist, *Structural insights into a yeast prion illuminate nucleation and strain diversity.* Nature, 2005. **435**(7043): p. 765-772.
37. Bryan, A.W., Jr., et al., *BETASCAN: probable beta-amyloids identified by pairwise probabilistic analysis.* PLoS Comput Biol, 2009. **5**(3): p. e1000333.
38. Berger, B., *Algorithms for protein structural motif recognition.* J Comput Biol, 1995. **2**(1): p. 125-38.
39. Baldwin, R.L., *In search of the energetic role of peptide hydrogen bonds.* J Biol Chem, 2003. **278**(20): p. 17581-8.
40. Menke, M., B. Berger, and L. Cowen, *Matt: Local Flexibility Aids Protein Multiple Structure Alignment.* PLoS Comput Biol, 2008. **4**(1): p. e10.
41. Waldispuhl, J., et al., *Modeling ensembles of transmembrane beta-barrel proteins.* Proteins, 2008. **71**(3): p. 1097-112.

42. Michelitsch, M. and J. Weissman, *A census of glutamine/asparagine-rich regions: Implications for their conserved function and the prediction of novel prions.* Proceedings of the National Academy of Sciences, 2000. **97**(22): p. 11910.
43. Alberti, S., et al., *A systematic survey identifies prions and illuminates sequence features of prionogenic proteins.* Cell, 2009. **137**(1): p. 146-58.
44. Krishnan, R. and S.L. Lindquist, *Structural insights into a yeast prion illuminate nucleation and strain diversity.* Nature, 2005. **435(7043)**: p. 765-72.
45. Bradley, P., et al., *BETAWRAP: successful prediction of parallel beta -helices from primary sequence reveals an association with many microbial pathogens.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(26): p. 14819-24.
46. McDonnell, A.V., et al., *Fold recognition and accurate sequence-structure alignment of sequences directing beta-sheet proteins.* Proteins, 2006. **63**(4): p. 976-85.
47. Petkova, A.T., et al., *A structural model for Alzheimer's beta -amyloid fibrils based on experimental constraints from solid state NMR.* Proc Natl Acad Sci U S A, 2002. **99**(26): p. 16742-7.
48. Sachse, C., M. Fandrich, and N. Grigorieff, *Paired beta-sheet structure of an Abeta(1-40) amyloid fibril revealed by electron microscopy.* Proc Natl Acad Sci U S A, 2008. **105**(21): p. 7462-6.
49. Zhang, C., J.L. Cornette, and C. Delisi, *Consistency in structural energetics of protein folding and peptide recognition.* Protein Sci, 1997. **6**(5): p. 1057-64.
50. Street, A.G. and S.L. Mayo, *Intrinsic beta-sheet propensities result from van der Waals interactions between side chains and the local backbone.* Proc Natl Acad Sci U S A, 1999. **96**(16): p. 9074-6.
51. Avbelj, F., P. Luo, and R.L. Baldwin, *Energetics of the interaction between water and the helical peptide group and its role in determining helix propensities.* Proc Natl Acad Sci U S A, 2000. **97**(20): p. 10786-91.
52. Pal, D. and P. Chakrabarti, *beta-sheet propensity and its correlation with parameters based on conformation.* Acta Crystallogr D Biol Crystallogr, 2000. **56**(Pt 5): p. 589-94.
53. Minor, D.L., Jr. and P.S. Kim, *Context is a major determinant of beta-sheet propensity.* Nature, 1994. **371**(6494): p. 264-7.
54. Gazit, E., *A possible role for pi-stacking in the self-assembly of amyloid fibrils.* FASEB J, 2002. **16**(1): p. 77-83.
55. Halfmann, R., et al., *Opposing effects of glutamine and asparagine govern prion formation by intrinsically disordered proteins.* Mol Cell, 2011. **43**(1): p. 72-84.
56. Jacobson, H. and W.H. Stockmayer, *Intramolecular Reaction in Polycondensations. I. The Theory of Linear Systems.* The Journal of Chemical Physics, 1950. **18**(12): p. 1600-1606.
57. Pace, C.N., et al., *Conformational stability and activity of ribonuclease T1 with zero, one, and two intact disulfide bonds.* J Biol Chem, 1988. **263**(24): p. 11820-5.
58. Petkova, A., et al., *Self-Propagating, Molecular-Level Polymorphism in Alzheimer's ß-Amyloid Fibrils*, in *Science*. 2005, American Association for the Advancement of Science. p. 262-265.
59. DePace, A.H., et al., *A critical role for amino-terminal glutamine/asparagine repeats in the formation and propagation of a yeast prion.* Cell, 1998. **93**(7): p. 1241-52.
60. Heise, H., et al., *Molecular-level secondary structure, polymorphism, and dynamics of full-length alpha-synuclein fibrils studied by solid-state NMR.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(44): p. 15871-6.
61. Sondheimer, N., et al., *The role of Sis1 in the maintenance of the [RNQ+] prion.* EMBO J, 2001. **20**(10): p. 2435-42.
62. Wickner, R.B., et al., *Prions beget prions: the [PIN+] mystery!* Trends Biochem Sci, 2001. **26**(12): p. 697-9.

Figure Legends

Figure 1, STITCHER results for *A*β (amyloid beta) *(m=1).* At left, a contact map of the 50 highest-scoring folds. The horizontal and vertical axes indicate, respectively, the residue numbers (counted from the N-terminus) of the lower and upper strands in each strand-pair of the structures. Starting locations of strand-pairs are indicated by circles, with size of circle (small to large) and color of circle (magenta to black, right-hand color spectrum) indicating the strength of the vote of the top 50 structures for that strand-pair. The strand-pairs are drawn along their length in shades of orange, with stability increasing from yellow to red (left-hand color spectrum). Fold structures are indicated by the dotted lines connecting strand-pairs into rungs and sheets. Structure scores are indicated by shades with stability increasing from blue to green (center color spectrum).
At right, the highest-scoring fold. Each strand-pair is denoted by its score (Sc) and its length (L). To the extent possible, rung-pairs proceed from left to right and sheets from top to bottom. Numbers to the left of the strands indicate the number of the residue immediately preceding the N-terminus of the strand. Slanted lines indicate the first residue of the strand, arrowheads the last residue of the strand, and connecting line(s) indicate the possible residue-pairing(s) of the first residues of the strands.

Figure 2, STITCHER results *(m=2)* for the two alleles of the *Podospora anserina* mating compatability protein: *A,* Het-s; *B,* Het-S. Contact maps, at left, and top-scoring structures, at right, are as described in the caption to Figure 1.

Figure 3, STITCHER results for Sup35p *(m=3)* in three species. *A, C. albicans*, *B, S. cerevisiae*, and *C, Y. lipolytica*. At left, a contact map of the 50 highest-scoring folds. At right, the top-scoring structure. For *S. cerevisiae*, the highest- scoring structures for $n = 2$ and $n = 3$ are presented. For description of colors, numbers, and lines, see the caption to Figure 1.

Figure 4, STITCHER results for, *A, S. cerevisiae* Rnq1p *(m=2),* and *B,* α-synuclein *(m=2).* At left, a contact map of the 50 highest-scoring folds. At right, a consensus structure assembled from all clusters of strand-pairs $\{(i_o - 2) \geq i \geq (i_o + 2), (j_o - 2) \geq j \geq (j_o + 2)\}$ found in $> 80\%$ of the top 50 highest-scoring folds. The range of strand positions and lengths is indicated by the shortest and longest possible strand arrows, drawn at the most N- and C-terminal possible locations. Connecting lines indicate the possible pairings of the first residues of the strands making up the strand-pair. The range of lengths and scores for a set of possible strand-pairs is indicated at L and Sc, respectively, for each set. The numbers to the left of the strands indicate the residue immediately before the leftmost possible residue in each strand (i.e., the residue before the first written above and below the strand arrows). For description of colors and numbers, see the caption to Figure 1.