Combinatorial Approaches to Probe the Sequence Determinants of Protein Aggregation and Amyloidogenicity

Christine Wurth[1,2], Woojin Kim[1] and Michael H. Hecht[1,*]

[1]*Department of Chemistry, Princeton University, Princeton, NJ 08544, USA;* [2]*Current Address: F. Hoffmann – La Roche AG, Pharma Division, CH-4070 Basel, Switzerland*

Abstract: Elucidation of the molecular determinants that drive proteins to aggregate is important both to advance our fundamental understanding of protein folding and misfolding, and as a step towards successful intervention in human disease. Combinatorial strategies enable unbiased and model-free approaches to probe sequence/structure relationships. Through the use of combinatorial methods, it is possible (i) to probe the sequence determinants of natural amyloid proteins by screening libraries of amino acid substitutions (mutations) to identify those that prevent amyloid formation; and (ii) to test new hypotheses about the mechanism of formation of amyloid fibrils by using these hypotheses to guide the design of combinatorial libraries of *de novo* amyloid-like proteins. Here, we review how these two approaches have been used to study the molecular determinants of protein aggregation and amyloidogenicity.

Keywords: Amyloid, combinatorial libraries, protein aggregation, protein design.

1. INTRODUCTION

The aggregation of soluble proteins into insoluble amyloid plaques is associated with a number of human pathologies including Alzheimer's disease, Type II diabetes, and Creuzfeld-Jakob disease [1-4]. In all these diseases, soluble proteins convert into structures that are rich in β-sheet and aggregate into highly insoluble amyloid fibrils. Although the amino acid sequences and native folds of the various amyloidogenic proteins are substantially different from one another, the amyloid fibrils themselves share similar structural features: All of them give rise to a characteristic cross-β X-ray fiber diffraction pattern.

Although recent studies suggest that polypeptides in general – irrespective of sequence – have some inherent propensity to assemble into amyloid-like fibrils [5], it is clear that some sequences are far more amyloidogenic than others. What are the molecular determinants that cause these amyloidogenic sequences to assemble into fibrils?

For well-folded proteins, the sequence determinants of structure and function are typically probed by two complementary approaches: structure determination and mutagenesis. High-resolution structural studies of amyloid fibrils are hampered by their insoluble, non-crystalline nature [6, 7]. Directed mutagenesis studies have been used to study amyloidogenic proteins [8-13]. However, by their very nature, such studies are usually directed to prove or disprove specific hypotheses about the role of particular residues or regions of a specific amyloidogenic sequence. Therefore, directed mutagenesis studies are rarely used as unbiased probes of the sequence determinants of amyloid formation.

*Address correspondence to this author at the Department of Chemistry, Princeton University, Princeton, NJ 08544, USA; Tel: +1-609+258-2901; Fax: +1-609-258-6746; Email: hecht@princeton.edu

Combinatorial methods represent a powerful approach to study systems for which our understanding of the underlying relationship between sequence and structure is incomplete. Combinatorial approaches have been used extensively to assess the tolerance of natural proteins to extensive amino acid substitutions [14-19]. Combinatorial methods have also enabled explorations of large regions of amino acid sequence space in searches for *de novo* sequences that are capable of folding into well folded and compact structures [20-26]. In recent years, combinatorial approaches have also been applied to amyloidogenic sequences. This review describes recent advances in applying combinatorial approaches and genetic screens to investigate the sequence determinants of protein aggregation and amyloidogenicity.

2. COMBINATORIAL APPROACHES TO PROBE AGGREGATION AND AMYLOIDOGENICITY

What are the molecular determinants of aggregation and amyloidogenicity? Which features of an amino acid sequence cause it to assemble into amyloid? Combinatorial methods can address these questions through two complementary approaches:

(i) A discovery driven approach can be used to probe the sequence determinants of natural amyloid proteins by screening libraries of amino acid substitutions (mutations) to identify those that prevent amyloid fibril formation.

(ii) A hypothesis driven approach can be used to test our understanding of sequence determinants by using this understanding to guide the rational design of combinatorial libraries of *de novo* amyloid proteins.

2.1. Combinatorial Libraries of Naturally-Occurring Amyloidogenic Sequences

An unbiased assessment of which features of a sequence are responsible for its propensity to aggregate can be ob-

0929-8665/06 $50.00+.00

tained by constructing and characterizing libraries of mutant sequences. Mutations that dramatically alter the aggregation behavior enable identification of the sequence determinants of aggregation for the wild-type protein.

Combinatorial libraries containing variants of naturally-occurring amyloidogenic sequences, such as the Alzheimer's Aβ sequence, can be constructed in a number of ways: In some cases, an entire protein sequence is targeted; in others, only a subset of a sequence is mutated. Moreover, in some cases a fully random mixture of all 20 amino acids is allowed; whereas in others, the mixture of amino acids is constrained to include only a subset of the 20 amino acids at any given position [26, 27].

One powerful method for constructing focused libraries of mutants is proline scanning mutagenesis. Because proline cannot be accommodated in standard α-helical or β-sheet secondary structures [28], proline scanning is especially useful for probing sequences wherein a stretch of α-helix or β-strand is essential to maintain the overall structure and/or function of a protein [29]. Since amyloid structure is known to be dominated by β-sheet structure, proline scanning mutagenesis is well-suited to assess which parts of a wild-type sequence are crucial for the formation of amyloid. This approach was applied to the Alzheimer's Aβ peptide by Morimoto *et al.*, who identified residues that are important in β-sheet formation in Aβ by systematically replacing each residue with proline at positions 19-26 of the full length Aβ(1-42) peptide [30]. All of the proline-substituted mutants except 22P and 23P showed a lower tendency to aggregate, and weaker cytotoxicity than wild type Aβ(1-42). Thus, the wild type Aβ(1-42) had an IC_{50} of 2.1 μM in an MTT (3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide) reduction assay on PC12 cells, whereas 19P, 21P, 24P and 26P had IC_{50}'s of 3.6, 100, 20 and 6.2 μM, respectively. Interestingly, mutant 22P aggregated to a greater extent, and displayed higher cytotoxicity than wild type Aβ(1-42) showing IC_{50} of 0.084 uM. Since proline has a propensity to occur in turns, the authors suggest that positions 22 and 23 form a turn in Aβ amyloid. Interestingly, position 22 is also the site of several naturally occurring mutations that lead to familial Alzheimer's disease. Examples include the E22Q (Dutch), E22K (Italian) and E22G (Arctic) mutations [31]. The early onset of Alzheimer's disease in these familial cases supports the importance of position 22 in the pathogenesis of the Aβ peptide.

Wetzel and coworkers used proline scanning mutagenesis initially to study fragments of the Aβ(1-40) peptide [32], and later to study the full sequence of the Aβ(1-40) peptide [33]. By substituting proline at each position in Aβ(1-40), these authors investigated the role of each residue in β-sheet formation and aggregation. Three segments in the sequence of Aβ(1-40) were found to be highly sensitive to proline replacement. These segments corresponded to three hydrophobic regions: residues 15-21 (the previously postulated central hydrophobic cluster for amyloid formation), residues 24-28, and residues 31-36. In contrast, residues 1-14, 22-23, 29-30, and 37-40 were rather insensitive to proline substitution. Based on the results of their proline scanning mutagenesis, Wetzel and coworkers proposed a model for the Aβ(1-40) protofilament. As shown in (Fig. **1**), the regions sensitive to

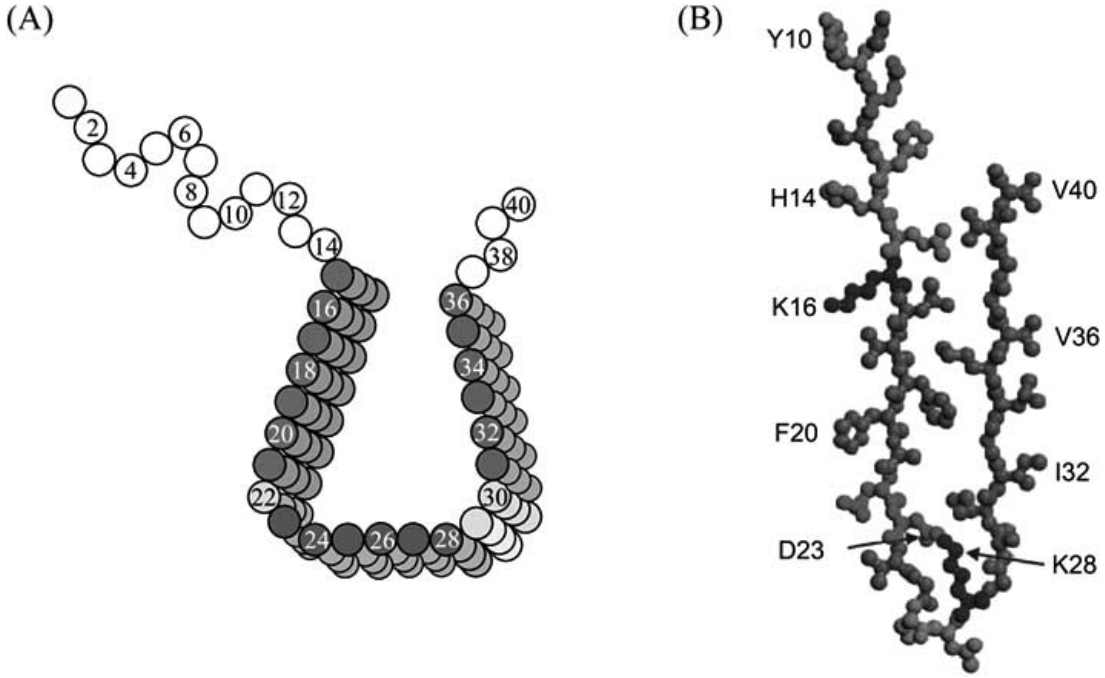

Figure 1. (A) Model of the Aβ(1-40) peptide protofilament based on scanning proline analysis. The chain of circles containing residue numbers represents one Aβ peptide molecule. Residues 1-14 and 37-40 are shown as disordered elements. Residues 15-21, 24-28, and 31-36 are shown as β-strands and residues 22, 23, 29 and 30 shown as turns. Residues 15-36 of additional Aβ peptide molecules are shown stacked in the H-bonding direction parallel to the fibril axis. From [33]. (B) Model of the structure of the Aβ(1-40) protofilament based on solid state NMR. Residues 1-9 are omitted. Residues 12-24 and 30-40 are shown as β-strands, and residues 25-29 are modeled as turns. D23 and K28 form salt bridge. From [34].

proline replacement are presumed to comprise the β-sheet portions of the fibrils. In contrast, the N-terminal 14 residues, which were insensitive to proline replacement are modeled to be relatively unstructured and are not included into the model of the amyloid fibril core. Residues 22-23 and 29-30 are proposed to be turn regions between the three segments of β-sheet structure.

Solid state NMR studies by Tycko and coworkers also show that the N-terminal 10 residues are relatively unstructured [34]. However, the model suggested by Tycko and coworkers has two β-sheet segments including residues 12-24 and 30-40. Residues 25-29 are in the turn region instead of forming another β-sheet segment.

While the Alzheimer's Aβ peptide is short enough to be suitable for systematic proline scanning of its entire sequence, longer amyloidogenic sequences are often studied by fragment-based approaches. For example, in studies of the Tau protein, von Bergen *et al.* identified a minimal hexapeptide motif, VQIVYK, which is capable of initiating the aggregation of Tau into pathological paired helical filaments (PHFs) [35]. This motif coincides with the sequence having the highest predicted β-sheet structure in the Tau sequence. These authors used proline scanning mutagenesis to investigate the importance of the hexapeptide during Tau aggregation. Systematic replacement of each residue in the minimal motif by proline prevented aggregation, supporting the hypothesis that the hexapeptide motif forms a local β-sheet structure within the Tau protein and initiates Tau aggregation.

Proline scanning mutagenesis was also used to study the sequence determinants of amyloidogenicity for human islet amyloid polypeptide (IAPP), the major protein component of fibril deposits associated with type II diabetes [36, 37]. The central region (residues 20-29) of IAPP is thought to play a key role in amyloid formation. Interestingly, comparison of the amyloidogenic human IAPP with the non-amyloidogenic rat sequence reveals that the rat sequence has three prolines among residues 20-29. To assess which residues of region 20-29 are most important in the aggregation of the human sequence, Moriarty and Raleigh synthesized the human IAPP (20-29) and a series of variants in which each position was replaced by a proline residue [38]. A proline substitution in the central region inhibited aggregation and amyloid formation, whereas substitutions in the peripheral regions did not have a dramatic effect.

Azriel and Gazit performed a systematic alanine scan of the short hexapeptide NFGAIL of IAPP [39]. This peptide corresponds to residues 23-28 of IAPP, and forms fibrils similar to those formed by the full-length 37 residue peptide. Alanine was chosen as the 'default' amino acid because it does not introduce polarity or charge, nor does it prevent β-sheet formation (as does proline) [40]. This work demonstrated that the substitution of the phenylalanine residue of NFGAIL to alanine led to a total loss of the ability of the peptide to assemble into amyloid fibrils. Why is this aromatic residue apparently so important for IAPP amyloid formation? Azriel and Gazit propose that π-stacking interactions play a significant role in the molecular recognition and self-assembly process by reducing the energetic barrier for the formation of amyloid fibrils. Phenylalanine residues were also found to play key roles in other amyloid-related sequences. For example the central hydrophobic cluster of the Aβ peptide (KLVFFA, residues 16-21), which has been shown to be essential for Aβ peptide fibril formation [41, 42], contains two phenylalanines, and several studies showed that replacement of Phe by non-aromatic residues decreases the ability of Aβ peptide to aggregate [32, 41-43].

Recently, our group probed the importance of phenylalanine versus other hydrophobic amino acids in the aggregation of Aβ(1-42) peptide. We constructed a combinatorial library of mutants in the full-length sequence of the Aβ(1-42) peptide in which all the nonpolar residues were randomized to other nonpolar residues – see below (Kim and Hecht, in preparation). Although large-scale substitution of nonpolar amino acids was generally tolerated (i.e. mutant peptides aggregated with properties similar to wild-type), a peptide lacking phenylalanine aggregated with much slower kinetics, especially during the nucleation step. These findings lend support to the proposal that aromatic residues play an important role in enhancing peptide aggregation.

An alternative strategy for probing the sequence determinants of aggregation was used by Salmona *et al.* in their studies of the prion protein responsible for Gerstmann-Straeussler-Scheinker (GSS) disease [44]. The major component of GSS amyloid is a PrP fragment spanning residues 82-146. Salmona *et al.* synthesized a peptide corresponding to residues 82-146. They also synthesized three variant peptides in which either (i) residues 106-126 were scrambled, (ii) residues 127-146 were scrambled, or (iii) the entire 82-146 sequence was scrambled. Aggregation studies revealed that the wild type peptide and the peptide in which residues 106-126 were scrambled both aggregated into protease resistant material. In contrast, peptides in which the C-terminal residues 127-146 were scrambled or in which the entire sequence was scrambled failed to aggregate. These results indicate that the integrity of the C-terminal region of this sequence is crucial for amyloid formation.

2.2. High Throughput Screens for Mutations that Modify Solubility and Aggregation

Combinatorial approaches are most powerful when they involve two key steps: first, a large and diverse library of molecules is generated; and second, the library is screened for sequences that possess desired properties [45-50]. Advances in molecular biology facilitate the generation of vast libraries of sequence variants [51, 52]. However, it is challenging to devise high-throughput assays that enable screening of millions of variants for protein folding and/or aggregation. This is particularly true for the amyloidogenic sequences, such as the Alzheimer's Aβ peptide, for which there is neither a selectable phenotype nor a screenable function associated with a 'correct' polypeptide fold.

In recent years, several new methods have been developed to screen for protein solubility (for review see [53]). Some of these methods use reporter tags fused to the sequence of interest. Other methods detect misfolded proteins *in vivo* using the bacterial stress response to protein misfolding. For example, Wigley *et al.* used the lacZα complementation assay as a solubility reporter system [54]. This assay is based on structural complementation between the

LacZα peptide fused to a test protein, and the inactive lacZω fragment of β-galactosidase: Activity of β-galactosidase is restored if the fusion protein remains soluble and the fused LacZα peptide is not hidden by aggregation. When LacZα was fused to the Alzheimer's Aβ peptide, inactive β-galactosidase was produced; however, a mutation that decreases Aβ peptide aggregation (F19P) restored β-galactosidase activity. The authors suggested that this screen could be used to identify drugs that inhibit fibril formation.

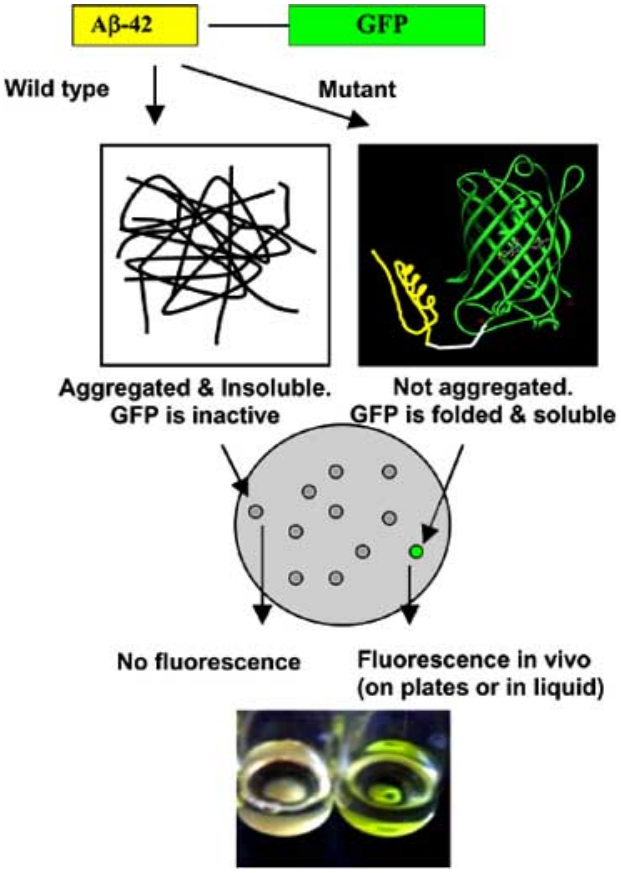

Figure 2. Schematic depiction of the properties of Aβ–GFP fusion proteins. The wild-type Aβ(1-42) peptide forms insoluble aggregates (left) and prevents the GFP portion of the fusion protein from forming its native fluorescent structure. However, mutations in the sequence of Aβ(1-42) that retard aggregation enable GFP to form its native green fluorescent structure (right). Structures of the Aβ(1-42) variants are unknown, and the yellow part of the ribbon diagram is merely a schematic cartoon. *E. coli* cells expressing GFP fusions to wild type (left) and mutant (right) forms of Aβ(1-42) appear white or green, respectively. The white/green color screen can be used either on plates (middle), or in liquid culture (bottom). Adapted from [43].

The availability of methods to screen for protein solubility has enabled experimentalists to probe the sequence determinants of amyloidogenicity by screening libraries of randomly generated amino acid substitutions (mutations) for those that prevent aggregation. Unlike studies of rationally designed mutations, screening of combinatorial libraries has the advantage of being model-independent and unbiased. We recently used this library-based approach to screen for soluble variants of the Aβ(1-42) peptide [43,72]. To isolate mutations that increase the solubility of the Aβ(1-42) peptide, we developed a genetic screen based on the folding of the reporter protein, GFP (green fluorescent protein) [55]. *E. coli* cells expressing fusions of the wild-type Aβ(1-42) peptide to GFP do not fluoresce, whereas randomly mutated variants of the Aβ(1-42) peptide with reduced tendencies to aggregate exhibited green fluorescence (Fig. **2**). The sequences of most of the variants with increased solubility support previous hypotheses emphasizing the importance of hydrophobic regions (such as the central hydrophobic cluster comprising residues 16-21) as determinants of Aβ peptide aggregation. The screen, however, also uncovered novel solubility-enhanced variants that were not predicted by existing models of Aβ amyloid formation. For example, the single mutation, Ala2Ser, and the double mutation, His6Leu + Gly38Asp, decrease aggregation. These N terminal and C terminal mutations are not predicted by existing models of Aβ fibril formation, which emphasize the central hydrophobic region. These results highlight the advantage of using combinatorial libraries and unbiased screens to elucidate the sequence determinants of aggregation and amyloidogenicity.

Whereas the GFP reporter system enables screens for mutations that enhance the solubility of amyloidogenic proteins, other systems have been developed to screen libraries for mutations that cause soluble and non-amyloidogenic proteins to become amyloidogenic. For example, Koscielska-Kasprzak and Otlewsky showed that novel sequences with high propensities to form amyloid can be found by screening libraries for sequences that are unusually resistant to proteolysis [56]. They displayed a 28-residue ββα zinc finger domain on the surface of phage, and selected for resistance to proteolytic degradation. Although their library was originally designed to find novel monomeric ββα peptides that are stable and well-folded in the absence of metal ions, the screen for proteolytic-resistant phage did not identify such ββα monomers. Instead, the selection yielded 8 peptides that were high in β-sheet content and formed aggregates. Three of these peptides were shown to self-assemble into amyloid-like fibrils. Why did a selection for resistance to proteolysis yield fibril-forming sequences? The authors suggest that hydrophobic and β-sheet rich peptides escaped proteolysis by interacting with neighboring phage domains rather than by forming an independent fold: Selection for protease resistance (unintentionally) yielded peptides with an overall hydrophobicity that was significantly higher than that of the natural zinc finger sequence. Furthermore, the design of the phage display library itself selected for β-sheet forming residues in β-strand regions and potential core positions. This resulted in a generally increased β-sheet content among the selected peptides. These results are consistent with the findings of Chiti *et al*, who showed that overall sequence hydrophobicity is a major determinant of amyloidogenicity [57].

2.3. Combinatorial Libraries of Amyloid-like Proteins Designed *De Novo*

In addition to the approaches described above, the molecular determinants of amyloidogenicity can also be probed by using combinatorial methods to design libraries of *de novo* amyloid proteins. As described above, screening libraries of substitutions in natural sequences enables an unbi-

ased search for the sequence determinants of aggregation. In contrast, the rational design of combinatorial libraries of *de novo* amyloid proteins enables a complementary approach wherein one can test particular hypotheses about possible sequence determinants by using these hypotheses as the basis for library design.

Protein design was used by Hecht and coworkers to probe the sequence determinants of amyloidogenicity. We constructed a combinatorial library of *de novo* sequences based on binary patterning of polar and nonpolar amino acids [58]. Our library was based on the alternating binary pattern O●O●O●O, where 'O' represents a combinatorial mixture of polar amino acids, and '●' represents a combinatorial mixture of nonpolar amino acids. As shown in (Fig. **3**), this pattern is consistent with the structural periodicity of an amphiphilic β-strand. Because amphiphilic β-strands are prone to aggregate [59-61], this alternating pattern might be expected to favor aggregation into amyloid fibrils. All sequences in the library were constrained by the O●O●O●O alternating pattern, however, the precise identities of the side chains were varied combinatorially: Polar residues were allowed to be His, Lys, Asn, Asp, Gln or Glu; and nonpolar residues were allowed to be Leu, Ile, Val or Phe. The *de novo* proteins were designed to fold into 6-stranded beta sheet structures, as shown schematically in (Fig. **3B**). The proteins were expressed from a library of synthetic genes and several were purified and characterized. Circular dichroism spectroscopy confirmed that the proteins indeed formed β-sheet structures; and electron microscopy demonstrated that they self-assembled into large oligomers resembling amyloid fibrils [58].

These libraries of *de novo* amyloid proteins can be compared with our work designing combinatorial libraries of α-helical proteins [22-24]. In both cases, libraries were designed by specifying the binary pattern of polar and nonpolar residues. Yet the resulting proteins displayed dramatically different properties: The α-helical sequences folded *intra*-molecularly into small globular domains. In contrast, the sequences with the alternating O●O●O●O pattern formed β-strands and self-assembled inter-molecularly into fibrillar structures. What causes these dramatically different structures? The determining difference is the binary patterning itself: In the earlier work, the library was constrained by the pattern O●OO●●OO●OO●O, consistent with the periodicity of amphiphilic α-helical structure. In contrast, the library constrained by the alternating O●O●O●O pattern comprised sequences consistent with the periodicity of amphiphilic β-strands. It is these amphiphilic β-strands that go on to aggregate into insoluble fibrils.

To probe whether the O●O●O●O pattern is inherently prone to favor aggregation, we asked whether disruption of this alternating pattern might prevent aggregation into fibrils. Toward this goal Wang and Hecht designed a series of mutants in which the central residue on the edge strand of each sheet (e.g. the first and last strands of six-stranded protein shown in (Fig. **3B**) was changed from a nonpolar residue to lysine [62]. In the redesigned β-strands, the binary pattern was changed from O●O●O●O to O●O**K**O●O (where **K** denotes lysine). The presence of a lysine on the nonpolar face of a β-strand should disfavor aggregated fibrils because such structures would bury an uncompensated charge. The nonpolar-to-lysine mutations, therefore, would be expected to favor monomeric structures in which the O●O**K**O●O sequences form the edge strands of β-sheets, with the charged lysine side chain accessible to solvent. Biophysical characterization of these redesigned proteins demonstrated that they indeed formed monomeric β-sheet proteins. This finding lends support to the hypothesis that uninterrupted stretches of alternating polar and nonpolar residues are inherently prone to aggregate into fibrils.

Recent work by DuBay *et al.* [63] further supports the hypothesis that alternating binary patterns favor aggregation. These authors developed an equation that predicts aggregation rate as a function of amino acid sequence. Three sequence-dependent factors were found to enhance the aggregation: These were hydrophobicity, lack of charge, and alternating patterns of polar and nonpolar amino acids [63].

If sequences with alternating patterns indeed have a high propensity to form amyloid fibrils, then one might expect such patterns to be disfavored by natural selection. To test this expectation, a database of 250,514 natural protein sequences was searched for all possible binary patterns. The search revealed that alternating patterns occur in nature significantly less often than other patterns with similar compositions [64]. The under-representation of alternating binary patterns in natural proteins, coupled with the observation that

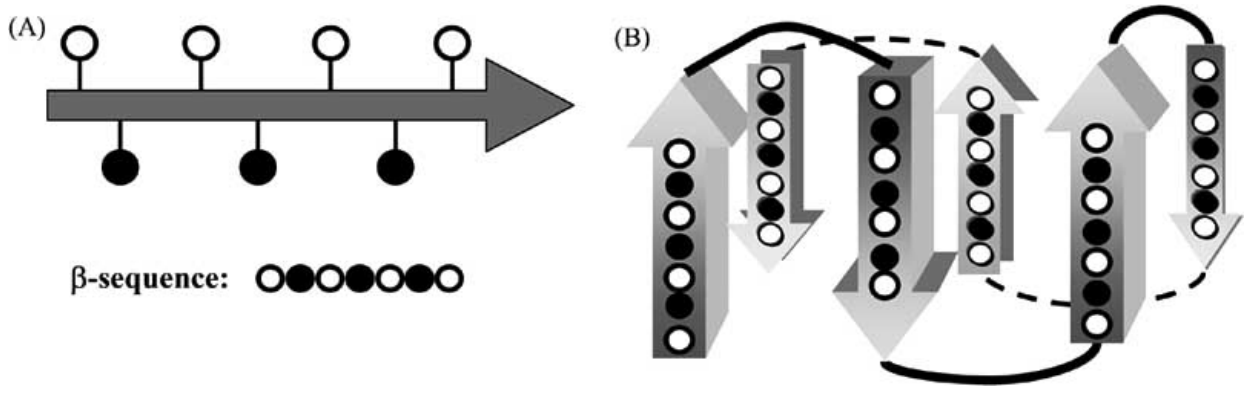

Figure 3. (A) Representation of an amphiphilic β-strand. Polar and nonpolar residues are shown in white and black, respectively. (B) Binary pattern for a library of six-stranded β-sheet proteins. Arrows designate β-strands.

such patterns promote amyloid formation in both natural and *de novo* proteins [58], suggests that sequences of alternating polar and nonpolar amino acids inherently favor amyloid, and consequently have been disfavored by evolutionary selection.

Alternating patterns are rare not only in the overall database of natural proteins but also in the database of naturally-occurring amyloidogenic sequences [58, 64] associated with pathologies including Alzheimer's disease, Type II diabetes, and Creuzfeld-Jakob disease. Although at first this may seem surprising, it is in fact the expected result, since these proteins were not selected by nature to be amyloidogenic: These amyloids are off-pathway misfolded structures of otherwise soluble, correctly folded proteins [5].

Although the binary code strategy was originally devised to enable the design of libraries of *de novo* proteins, it has also been used recently to probe the sequence determinants of aggregation in a natural system. Recently, Kim and Hecht (unpublished) used the binary code to study the role of sequence hydrophobicity in the aggregation of the Alzheimer's Aβ(1-42) peptide. In this study, 12 nonpolar residues in the four hydrophobic stretches of Aβ(1-42) were mutated randomly to a combinatorial mixture of nonpolar residues. The resulting library was screened *in vivo* for aggregation versus solubility using the GFP reporter described above. The experiments revealed that these Aβ variants still aggregated – even though the 42 residue sequence was mutated in 8 to 12 positions. Several of the most heavily mutated peptides were synthesized (in the absence of GFP) and characterized. Aggregation studies and fluorescence measurements using Thioflavin-T confirmed that despite the large number of mutations, many of these variants were similar to the wild-type sequence in their abilities to form large aggregates dominated by β-sheet secondary structure. These findings suggest that – at least in some cases – the hydrophobic patterning of a sequence has a greater impact on aggregation than the exact identities of the nonpolar side chains.

2.4. Simplified Models of Amyloidogenic Sequences

The size and/or complexity of natural amyloidogenic sequences has prompted several groups to search for simplified model sequences to elucidate the process of amyloid formation [65]. Lopez de la Paz *et al.* used computational methods to search for hexapeptide sequences optimized to self-associate into homopolymeric β-sheet structures [66]. They assumed that propagation and stacking of preformed β-sheets would lead to the assembly of amyloid fibrils. Sequences suggested by their computational studies were then synthesized and characterized. The experiments revealed that the peptides predicted to form polymeric β-sheets only formed well-defined fibrils if the total net charge of the molecule was ± 1. Neutral or more highly charged sequences were less prone to form fibrils. Neutral peptides formed amorphous aggregates rather than fibrils, whereas more highly charged peptides were prevented from forming fibrils by repulsion of excess uncompensated charge. These findings suggest that both the conformational state of the polypeptide and the net charge play roles in amyloid fibril formation, and that in more complex protein systems, the formation of amyloid fibrils might occur preferentially for particular charge states of the macromolecule.

In more recent work, Lopez de la Paz and Serrano synthesized a combinatorial library of related sequences by performing positional scanning mutagenesis on the designed hexapeptide, STVIIE [67]. They systematically replaced all residues in STVIIE with all the natural amino acids except Cys. This study revealed that both restrictive (sequence core), and tolerant (sequence periphery) mutations could be found within this short amyloidogenic sequence. The amyloidogenicity of the hexapeptide was determined mainly by core residues; the peripheral residues acted primarily as amyloid modulators. This finding is consistent with that of Moriarty and Raleigh [38], who found that in the IAPP amyloidogenic region 20-29, a Pro substitution was tolerated in the periphery, but not in the core sequence (see above). The residues VII were found to be highly amyloidogenic, but aggregation could be avoided if this motif was surrounded by an excess of charged residues, which enhances solubility, and/or proline, which disrupts β-structure. Charged residues were allowed only at the ends of the sequence. Phe was the only amino acid that allowed fibril formation at any position, consistent with the work of Azriel and Gazit, who proposed that π-stacking of aromatic residues favors self-assembly into amyloid fibrils (see above) [39]. From their work, Lopez de la Paz and Serrano were able to specify a sequence pattern that facilitates the identification of potentially amyloidogenic sequences in proteins [67]. The predictive power of this pattern was validated by combinatorial experiments *in vitro*. Moreover, *in silico* sequence scanning of amyloid proteins also supported the pattern: for example residues 16-21 of the Aβ peptide, which have long been considered to be important for amyloid formation, are consistent with the newly described pattern.

A variety of globular proteins have been shown to form amyloid fibrils *in vitro* under selected conditions, especially when the protein structure is partially unfolded [68]. Fortunately, however, under physiological conditions amyloid formation *in vivo* seems to be limited to a small number of proteins and peptides. What distinguishes this small number of proteins that form amyloid under physiological conditions from those that do not? Tjernberg *et al.* showed that peptides as short as 4 residues with inherent propensity to form beta-strand conformation can form amyloid fibrils practically identical to fibrils associated with amyloid diseases [69]. They suggest this finding may explain why attempts to find common motifs among amyloidogenic proteins have so far been largely unsuccessful. Natural beta-strand proteins often avoid aggregation by protecting their edge strands with polar or irregular stretches [70]. Thus the determinants of amyloidogenicity seem to be found not only in the 'amyloidogenic region': Equally important are the regions surrounding (and protecting) an amyloidogenic stretch.

3. CONCLUSIONS

Combinatorial approaches enable powerful studies of amyloidogenic polypeptide systems for which our understanding of the relationship between sequence and structure is incomplete. Through the use of combinatorial methods it is possible to probe the sequence determinants of natural amyloid proteins by screening mutant libraries to identify

those substitutions that prevent amyloid; and to test hypotheses about amyloidogenesis by constructing combinatorial libraries of *de novo* amyloid peptides and proteins. In recent years, the use of combinatorial methods has led to the identification of a range of sequence determinants of amyloidogenecity including (i) the β-sheet propensity of the amino acids; (ii) the overall hydrophobicity of the sequence; (iii) the net charge of the molecule and the position of the charged residues; (iv) the role of the peripheral regions of an amyloidogenic sequence; (v) the effect of individual residues like prolines and aromatic amino acids (π-stacking); and (vi) the binary patterning of polar and nonpolar residues.

Combinatorial experiments on natural and designed sequences have provided a large database of sequences that display a wide range of aggregation properties. Analysis of the correlations between these sequences and the resulting properties has enabled researchers to devise algorithms that predict with reasonable accuracy the propensity of a peptide or protein to aggregate into amyloid-like structures [63, 71].

Future combinatorial studies will advance our fundamental understanding of protein folding and misfolding, will contribute further to the elucidation of the molecular determinants of amyloidogenicity, and will thereby provide a rational basis for successful intervention in human amyloid diseases.

ABBREVIATIONS

IAPP = Islet amyloid polypeptide

GFP = Green fluorescent protein

GSS = Gerstmann-Straeussler-Scheinker

MTT = 3-(4,5-Dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide

PHFs = Paired helical filaments.

REFERENCES

[1] Kelly, J. **(1996)** *Curr. Opin. Struct. Biol., 6*, 11-17.
[2] Prusiner, S.B. **(1997)** *Science, 278*, 245-251.
[3] Selkoe, D. **(2001)** *Physiol. Rev., 81*, 741-766.
[4] Dobson, C.M. **(1999)** *Trends Biochem. Sci., 24*, 329-332.
[5] Dobson, C.M. **(2003)** *Nature, 426*, 884-890.
[6] Serpell, L.C. **(2000)** *Biochim. Biophys. Acta, 1502*, 16-30.
[7] Lynn, D. and Meredith, S. **(2000)** *J. Struct. Biol., 130*, 153-173.
[8] Hammarström, P., Jiang, X., Hurshman, A., Powers, E. and Kelly, J. **(2002)** *Proc. Natl. Acad. Sci. USA, 99*, 16427-16432.
[9] Villegas, V., Zurdo, J., Filimonov, V.V., Aviles, F.X., Dobson, C.M. and Serrano, L. **(2000)** *Protein Sci., 9*, 1700-1708.
[10] Esler, W., Stimson, E., Ghilardi, J., Lu, Y., Felix, A., Vinters, H., Mantyh, P., Lee, J. and Maggio, J. **(1996)** *Biochemistry, 35*, 13914-13921.
[11] Fay, D., Fluet, A., Johnson, C. and Link, C. **(1998)** *J. Neurochem., 71*, 1616-1625.
[12] Fraser, P., McLachlan, D., Surewicz, W., Mizzen, C., Snow, A., Nguyen, J. and Kirschner, D. **(1994)** *J. Mol. Biol., 244*, 64-73.
[13] Päiviö, A., Nordling, E., Kallberg, Y., Thyberg, J. and Johansson, J. **(2004)** *Protein Sci., 13*, 1251-1259.
[14] Lim, W. and Sauer, R. **(1989)** *Nature 339*, 31-36.
[15] Lim, W. and Sauer, R. **(1991)** *J. Mol. Biol., 219*, 359-376.
[16] Axe, D., Foster, N. and Fersht, A. **(1996)** *Proc. Natl. Acad. Sci. USA, 93*, 5590-5594.
[17] Gu, H., Yi, Q., Bray, S., Riddle, D., Shiau, A. and Baker, D. **(1995)** *Protein Sci., 4*, 1108-1117.
[18] Munson, M., O'Brien, R., Sturtevant, J. and Regan, L. **(1994)** *Protein Sci., 3*, 2015-2022.
[19] Palzkill, T. and Botstein, D. **(1992)** *Proteins, 14*, 29-44.
[20] Mandecki, W. **(1990)** *Protein Eng., 3*, 221-226.
[21] Davidson, A. and Sauer, R. **(1994)** *Proc. Natl. Acad. Sci. USA, 91*, 2146-2150.
[22] Kamtekar, S., Schiffer, J., Xiong, H., Babik, J. and Hecht, M. **(1993)** *Science, 262*, 1680-1685.
[23] Wei, Y., Liu, T., Sazinsky, S., Moffet, D., Pelczer, I. and Hecht, M. **(2003a)** *Protein Sci., 12*, 92-102.
[24] Wei, Y., Kim, S., Fela, D., Baum, J. and Hecht, M. **(2003b)** *Proc. Natl. Acad. Sci. USA, 100*, 13270-13273.
[25] Ventura, S. and Serrano, L. **(2004)** *Proteins, 56*, 1-10.
[26] Hecht, M., Das, A., Go, A., Bradley, L. and Wei, Y. **(2004)** *Protein Sci., 13*, 1711-1723.
[27] Lahr, S., Broadwater, A., Carter, C., Collier, M., Hensley, L., Waldner, J., Pielak, G. and Edgel, M. **(1999)** *Proc. Natl. Acad. Sci. USA, 96*, 14860-14865.
[28] Creighton, T. **(1992)** in *Proteins* 2nd Ed. pp.171-199. W.H. Freeman and Company, New York.
[29] Schulman, B. and Kim, P. **(1996)** *Nat. Struct. Biol., 3*, 682-687.
[30] Morimoto, A., Irie, K., Murakami, K., Ohigashi, H., Shindo, M., Nagao, M., Shimizu, T. and Shirasawa, T. **(2002)** *Biochem. Biophys. Res. Com., 295*, 306-311.
[31] Murakami, K., Irie, K., Morimoto, A., Ohigashi, H., Shindo, M., Nagao, M., Shimizu, T. and Shirasawa, T. **(2002)** *Biochem. Biophys. Res. Com., 294*, 5-10.
[32] Wood, S.J., Wetzel, R., Martin, J.D. and Hurle, M.R. **(1995)** *Biochemistry, 34*, 724-730.
[33] Williams, A.D., Portelius, E., Kheterpal, I., Guo, J., Cook, K.D., Xu, Y. and Wetzel, R. **(2004)** *J. Mol. Biol., 335*, 833-842.
[34] Petkova, A., Ishii, Y., Balbach, J., Antzutkin, O., Leapman, R., Delaglio, F. and Tycko, R. **(2002)** *Proc. Natl. Acad. Sci. USA, 99*, 16742-16747.
[35] von Bergen, M., Friedhoff, P., Biernat, J., Heberle, J., Mandelkow, E.M. and Mandelkow, E. **(2000)** *Proc. Natl. Acad. Sci. USA, 97*, 5129-5134.
[36] Westermark, P., Wernstedt, C., Wilander, E., Hayden, D., O'Brien, T. and Johnson, K. **(1987)** *Proc. Natl. Acad. Sci. USA, 84*, 3881-3885.
[37] Cooper, G., Willis, A., Clark, A., Turner, R., Sim, R. and Reid, K. **(1987)** *Proc. Natl. Acad. Sci. USA, 84*, 8628-8632.
[38] Moriarty, D. and Raleigh, D. **(1999)** *Biochemistry, 38*, 1811-1818.
[39] Azriel, R. and Gazit, E. **(2001)** *J. Biol. Chem., 276*, 34156-34161.
[40] Cunningham, B. and Wells, J. **(1989)** *Science, 244*, 1081-1085.
[41] Hilbich, C., Kisters-Woike, B., Reed, J., Masters, C. and Beyreuther, K. **(1992)** *J. Mol. Biol.*, 228, *460*-473.
[42] Tjernberg, L.O., Näslund, J., Lindqvist, F., Johansson, J., Karlström, A., Thyberg, J., Terenius, L. and Nordstedt, C. **(1996)** *J. Biol. Chem., 271*, 8545-8548.
[43] Wurth, C., Guimard, N.K. and Hecht, M.H. **(2002)** *J. Mol. Biol., 319*, 1279-1290.
[44] Salmona, M., Morbin, M., Massignan, T., Colombo, L., Mazzoleni, G., Capobianco, R., Diomede, L., Thaler, F., Mollica, L., Musco, G., Kourie, J.J., Bugiani, O., Sharma, D., Inouye, H., Kirschner, D.A., Forloni, G. and Tagliavini, F. **(2003)** *J. Biol. Chem., 278*, 48146-48153.
[45] Roy, S., Helmer, K.J. and Hecht, M.H. **(1997)** *Fold Des., 2*, 89-92.
[46] Rosenbaum, D.M., Roy, S. and Hecht, M.H. **(1999)** *J. Am. Chem. Soc., 121*, 9509-9513.
[47] Keefe, A. and Szostak, J. **(2001)** *Nature, 410*, 715-718.
[48] Matsuura, T., Ernst, A. and Plückthun, A. **(2002)** *Protein Sci., 11*, 2631-2643.
[49] Matsuura, T., Ernst, A., Zechel, D. and Plückthun, A. **(2004)** *Chem. Biochem. 5*, 177-182.
[50] Lin, H. and Cornish, V. **(2002)** *Angew. Chemie Int. Ed., 41*, 4402-4425.
[51] Zaccolo, M., Williams, D., Brown, D. and Gherardi, E. **(1996)** *J. Mol. Biol., 255*, 589-603.
[52] Kunichika, K., Hashimoto, Y. and Imoto, T. **(2002)** *Protein Eng., 15*, 805-809.
[53] Waldo, G.S. **(2003)** *Curr. Opin. Chem. Biol., 7*, 33-38.
[54] Wigley, W.C., Stidham, R.D., Smith, N.M., Hunt, J.F. and Thomas, P.J. **(2001)** *Nat. Biotech., 19*, 131-136.
[55] Waldo, G.S., Standish, B.M., Berendzen, J. and Terwilliger, T.C. **(1999)** *Nat. Biotech., 17*, 691-695.
[56] Koscielska-Kasprzak, K. and Otlewski, J. **(2003)** *Protein Sci., 12*, 1675-1685.

[57] Chiti, F., Stefani, M., Taddei, N., Ramponi, G. and Dobson, C.M. **(2003)** *Nature, 424*, 805-808.
[58] West, M.W., Wang, W., Patterson, J., Mancias, J.D., Beasley, J.R. and Hecht, M.H. **(1999)** *Proc. Natl. Acad. Sci. USA, 96*, 11211-11216.
[59] Xiong, H., Buckwalter, B., Shieh, H. and Hecht, M. **(1995)** *Proc. Natl. Acad. Sci. USA, 92*, 6349-6353.
[60] Brack, A. and Spach, G. **(1981)** *J. Am. Chem. Soc., 103*, 6319-6323.
[61] Zhang, S., Holmes, T., Lockshin, C. and Rich, A. **(1993)** *Proc. Natl. Acad. Sci., USA* 90, 3334-3338.
[62] Wang, W. and Hecht, M.H. **(2002)** *Proc. Natl. Acad. Sci. USA, 99*, 2760-2765.
[63] DuBay, K.F., Pawar, A.P., Chiti, F., Zurdo, J., Dobson, C.M. and Vendruscolo, M. **(2004)** *J. Mol. Biol., 341*, 1317-1326.
[64] Broome, B. and Hecht, M. **(2000)** *J. Mol. Biol., 296*, 961-968.
[65] Kammerer, R., Kostrewa, D., Zurdo, J., Detken, A., Garcia-Echeverria, C., Green, J.D., Müller, S.A., Meier, B.H., Dobson, C.M. and Steinmetz, M. **(2003)** *Proc. Natl. Acad. Sci. USA, 101*, 4435-4440.
[66] Lopez de la Paz, M., Goldie, K., Zurdo, J., Lacroix, E., Dobson, C.M., Hoenger, A. and Serrano, L. **(2002)** *Proc. Natl. Acad. Sci. USA, 99*, 16052-16057.
[67] Lopez de la Paz, M. and Serrano, L. **(2004)** *Proc. Natl. Acad. Sci. USA, 101*, 87-92.
[68] Fändrich, M., Fletcher, M. and Dobson, C.M. **(2001)** *Nature, 410*, 165-166.
[69] Tjernberg, L., Hosia, W., Bark, N., Thyberg, J. and Johansson, J. **(2002)** *J. Biol. Chem., 277*, 43243-43246.
[70] Richardson, J.S. and Richardson, D.C. **(2002)** *Proc. Natl. Acad. Sci. USA, 99*, 2754-2759.
[71] Fernandez-Escamilla, A., Rousseau, F., Schymkowitz, J. and Serrano, L. **(2004)** *Nat. Biotech., 22*, 1302-1306.
[72] Kim, W. and Hecht, M.H. (**2005**) *J. Biol. Chem., 280*, 35069-35076.

Received: December 15, 2004 Revised: April 07, 2005 Accepted: April 19, 2005