Sequence determinants of amyloid fibril formation

Manuela López de la Paz* and Luis Serrano

European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany

Edited by Peter G. Wolynes, University of California at San Diego, La Jolla, CA, and approved October 15, 2003 (received for review August 1, 2003)

The establishment of rules that link sequence and amyloid feature is critical for our understanding of misfolding diseases. To this end, we have performed a saturation mutagenesis analysis on the *de novo*-designed amyloid peptide STVIIE (1). The positional scanning mutagenesis has revealed that there is a position dependence on mutation of amyloid fibril formation and that both very tolerant and restrictive positions to mutation can be found within an amyloid sequence. In this system, mutations that accelerate β-sheet polymerization do not always lead to an increase of amyloid products. On the contrary, abundant fibrils are typically found for mutants that polymerize slowly. From these experiments, we have extracted a sequence pattern to identify amyloidogenic stretches in proteins. The pattern has been validated experimentally. *In silico* sequence scanning of amyloid proteins also supports the pattern. Analysis of protein databases has shown that highly amyloidogenic sequences matching the pattern are less frequent in proteins than innocuous amino acid combinations and that, if present, they are surrounded by amino acids that disrupt their aggregating capability (amyloid breakers). This study provides the potential for a proteome-wide scanning to detect fibril-forming regions in proteins, from which molecules can be designed to prevent and/or disrupt this process.

It is accepted that misfolding of peptides and proteins cause the fibrillar aggregates that characterize the group of diseases known as amyloidoses (1, 2). Despite active research, a detailed understanding of the molecular principles underlying the transformation of soluble proteins into amyloid aggregates is still lacking (1–3).

The group of peptides and proteins capable of forming amyloid fibrils is very diverse (4–8). This group does not only consist of proteins involved in amyloid deposits *in vivo* (4), but also of nonpathogenic (15, 16) and designed peptides and proteins (7, 8). X-ray fiber diffraction data indicate that all amyloid fibrils share a cross-β-structure, regardless of the sequence or native fold of the soluble precursor (1, 2). Thus, an increasingly adopted view is that the ability to form amyloid fibrils is a general property of the polypeptide backbone (6). Nevertheless, the propensity of a given polypeptide to form amyloid fibrils depends enormously on amino acid composition (7–10).

A protein has to be partially or fully unfolded to aggregate (1, 2). Yet, mutations leading to aggregation can alter (9, 10), or not alter (11, 12), protein stability. Aggregation from some peptides and proteins involved in relevant amyloidoses, such as type II diabetes and Alzheimer's and Parkinson's disease, does not require, however, previous unfolding, because these molecules are largely unstructured under physiological conditions (2). Nevertheless, most of the natively unfolded proteins *in vivo* do not undergo aggregation (13), indicating that unfolding is necessary, but not sufficient, to promote aggregation. Hence, there must be some sequence motifs that, once they become exposed, are more prone to aggregation than others. In fact, experimental evidence is compelling in favor of the hypothesis that small regions of a protein are responsible for its amyloidogenic behavior (14, 15). If amyloid aggregation is actually driven by short fragments of a misfolded protein, small-model peptides should be more suitable than proteins to investigate those elements in sequences that favor aggregation. Whereas a mutation would alter just the self-assembly properties of a small peptide, it might lead to protein destabilization, complicating the extraction of pure sequence propensities to form amyloid fibrils.

We have recently reported the computer-aided design of a peptide-based model system for amyloidogenesis (8). Its simplicity has served to highlight that fibril formation is due to a very delicate balance between specific side chain and electrostatic interactions within a sequence. Now, we have exploited the small size of these peptides to determine how exact sequence details modulate, or completely disrupt, the apparent generality of amyloid fibril formation in proteins (1, 6). This question has been addressed by systematically replacing the residues of the designed amyloid peptide STVIIE (**1**) with all natural amino acids, except Cys. Although some Pro and Ala scannings on amyloid sequences have been published (15, 16), the mutagenesis experiment presented here constitutes, to our knowledge, the only full positional scanning performed so far. From these experiments, we have derived an amyloid sequence pattern for the identification of potential amyloidogenic regions in proteins. In the light of the knowledge acquired in this study, we believe that, in same way that it has been shown for globular proteins, there are general rules governing the amyloidogenicity of a polypeptide chain.

Methods

Peptide Synthesis and Purification. Peptides were synthesized by using the standard fluorenylmethoxycarbonyl solid-phase chemistry. Peptide homogeneity and composition were analyzed by analytical HPLC and MS (85–95% purity). The "X" positions of the peptide mixtures $X_1X_2V_3I_4I_5X_6$, $X_1X_2L_3N_4F_5X_6$, $X_1X_2L_3E_4F_5X_6$, and $X_1X_2W_3E_4F_5X_6$ were incorporated by coupling a mixture of 13 L-amino acids (Ala, Met, Phe, Val, Ile, Leu, Thr, Tyr, Asn, Gln, Glu, Asp, and Gly) with the relative ratio suitability adjusted to yield close to equimolar incorporation. The quality of the peptide mixtures was validated by electrospray MS.

Fibril Sample Preparation. Fibril samples were prepared as reported (8). Samples were incubated at room temperature and checked by CD ($t = 0$, 1 month) and electron microscopy (EM) ($t = 1$ month). After an incubation time of 20 d, 1-mM solutions of the peptide mixtures were concentrated by ultracentrifugation at $300{,}000 \times g$ for 1 h and the resulting pellet was resuspended in buffer (final concentration ≈10 mM).

Far-UV CD and EM. Far-UV CD and EM were carried out as reported (8).

Sequence Scanning of Amyloid Proteins. Sequence patterns for the identification of six-residue amyloid fragments in proteins were based on the allowed amino acids substitutions at each position of **1** (see *Results and Discussion*).

Cys was not used in the mutagenesis experiment. However, due to its similarity to Ser, we assumed that is allowed (or forbidden) at the same positions as Ser. Protein sequences were scanned by using PATTINPROT, which can be accessed at http://npsa-pbil.ibcp.fr (17).

Sequence Scanning of Protein Databases. Observed values of a given sequence motif X (O_X) were obtained by scanning protein data-

This paper was submitted directly (Track II) to the PNAS office.

Abbreviation: EM, electron microscopy.

*To whom correspondence should be addressed. E-mail: delapaz@embl.de.

Table 1. Positional scanning mutagenesis on the amyloid peptide 1 (STVIIE): Estimation of the β-sheet population in solution at *t* = 0 by CD [$\Theta_{217}/\Theta_{208}$ (percent β)] and EM fibril detection at *t* = 1 month

	X_1-TVIIE		S-X_2-VIIE		ST-X_3-IIE		STV-X_4-IE		STVI-X_5-E		STVII-X_6	
Amino acid substitution*	$\Theta_{217}/\Theta_{208}$ (percent β)	EM	$\Theta_{217}/\Theta_{208}$ (percent β)	EM	$\Theta_{217}/\Theta_{208}$ (percent β)	EM	$\Theta_{217}/\Theta_{208}$ (percent β)	EM	$\Theta_{217}/\Theta_{208}$ (percent β)	EM	$\Theta_{217}/\Theta_{208}$ (percent β)	EM
Gly (+1)	+0.67 (17)	+++	+0.78 (20)	+	+0.61 (16)	−	+0.49 (13)	−	+0.37 (10)	−	+*2.56 (66)*	+++[†]
Ala (+1)	+0.69 (18)	+++	−*1.63 (100)*	++	+0.42 (11)	−	+0.45 (12)	−	+0.45 (12)	−	+*2.46 (63)*	+++
Val (+1)	−*2.38 (100)*	+	+1.10 (28)	++	+**0.68 (17)**	+++	+0.75 (19)	−	+0.65 (17)	−	+*2.44 (62)*	+++
Ile (+1)	−*1.29 (100)*	++	−*98.7 (100)*	++	+1.19 (30)	−	+**0.68 (17)**	+++	+**0.68 (17)**	+++	+*1.46*[‡] *(37)*	++
Leu (+1)	+*3.20 (82)*	++	+*2.42 (62)*	++	+0.62 (16)	+++	+1.16 (30)	+++	+0.56 (14)	−	+*5.47 (100)*	+
Met (+1)	+1.14 (29)	+	−*1.12 (100)*	++	+0.45 (12)	−	+0.46 (12)	−	+0.47 (12)	−	−*8.10 (100)*	+++
Ser (+1)	+**0.68 (17)**	+++	+0.73 (19)	++	+0.46 (12)	+	+0.50 (13)	−	+0.54 (14)	−	+0.85 (22)	++
Thr (+1)	+1.41 (36)	++	+**0.68 (17)**	+++	+0.54 (14)	−	+0.63 (16)	++	+0.58 (15)	−	−*1.62 (100)*	+++
Tyr (+1)	−*1.88 (100)*	+	+1.08 (28)	++	+0.93 (24)	−	+1.84 (47)	++	+0.70 (18)	++	+1.12 (29)	++
Trp (+1)	+*3.11 (80)*	++	+0.39 (10)	−	+1.38 (35)	++[†]	+0.52 (13)	++	+0.50 (13)	−	+0.76 (19)	++
Phe (+1)	+0.62 (16)	++	+*3.70 (95)*	+	−*9.01 (100)*	++	−*0.05 (100)*	+++	+0.32 (8)	+++	+*3.93 (100)*	+++
Asn (+1)	+*7.41 (100)*	++	+*1.3*[‡] *(34)*	+	−*1.38 (100)*	++	+2.27 (58)	+++	+0.68 (17)	−	+*3.61 (92)*	++
Asp^0 (+1)	+0.92 (23)	++	+0.98 (25)	++	+0.64 (16)	−	+1.09 (28)	−	+0.67 (17)	−	+*1.00*[‡] *(26)*	++
Asp$^-$ (−1)	+0.71 (18)	++	+0.63 (16)	+	+0.59 (15)	−	+0.57 (15)	−	+0.58 (15)	−	+0.70 (18)	++
Gln (+1)	+*19.90 (100)*	+	+0.77 (20)	+++	+*46.49 (100)*	+	+0.41 (11)	−	+0.39 (10)	−	+*2.06 (53)*	+++
Glu^0 (+1)	−*1.25 (100)*	++	+0.75 (19)	+++	+0.96 (25)	+	+0.60 (15)	+++	+0.48 (12)	−	+**0.68 (17)**	+++
Glu$^-$ (−1)	+*1.14*[‡] *(29)*	+++	+0.69 (18)	+	+0.59 (15)	−	+0.51 (13)	−	+0.59 (15)	−	+0.70 (18)	+
Lys$^+$ (+1)	+0.75 (19)	+++	+0.66 (17)	−	+0.58 (15)	−	+0.54 (14)	−	+0.54 (14)	−	+0.69 (18)	−
Arg$^+$ (+1)	+*1.39*[‡] *(36)*	+++	+0.69 (18)	−	+0.61 (16)	−	+0.58 (15)	−	+0.60 (15)	−	+0.72 (18)	−
His$^+$ (+1)	+0.97 (25)	+++	+0.67 (17)	−	+0.55 (14)	−	+0.50 (13)	−	+0.53 (14)	−	+0.69 (18)	−
His^0 (−1)	+0.94 (24)	++	+0.60 (15)	−	+0.57 (15)	−	+0.50 (13)	−	+0.52 (13)	−	+0.63 (16)	−
Pro (+1)	+0.73 (19)	−	+0.64 (16)	−	+0.65 (17)	−	+0.52 (13)	−	+0.62 (16)	−	+0.47 (12)	−

To compare the β-sheet content (percent β) of the peptides, we have calculated the ratio between the ellipticity at 217 nm (Θ_{217}), β-sheet minimum, and the ellipticity at 208 nm (Θ_{208}), isodichroic point of the random coil→β-sheet transition (ratio $\Theta_{217}/\Theta_{208}$). This ratio is concentration-independent (39). The percentage of β-sheet population has been estimated assuming that poly-L-lysine is a good approximation of the 100% of β-sheet conformation for polymeric β-sheets (40). Poly-L-lysine in a β-sheet conformation shows an ellipticity of −18.400 deg·cm^2·dmol^{-1} at 217 nm and of −4.700 at 208 (ratio $\Theta_{217}/\Theta_{208} = +3.91$) (40). Ratios higher than +3.91 and negative ratios are indicative of a percentage of β-structure of ~100%. EM, the amount of fibrils in a sample has been quantified by visual inspection of the grids: +++, large amount of fibrils; ++, medium amount of fibrils; +, scarce amount of fibrils; −, no fibrils at all. Data from 1 are bold. Italicized annotations refer to amino acid substitutions that lead to an acceleration of β-sheet polymerization (percent β ≥ 50% at *t* = 0) with respect to **1**. Data from replacements leading to a high amyloid feature (+++), but a slow polymerization (percent β ≤ 50% at *t* = 0) are underlined.

*The termini of the peptides and the pH of the solutions have been set up accordingly to induce fibril formation (net change = ±1) (8). Mutants generated by replacement with a noncharged amino acid have been synthesized with free termini (pH 2.6, net charge of +1). If positively charged residues are used, then molecules have both termini-protected (pH 2.6, net charge of +1). Asp and Glu substitutions are incorporated into molecules with free termini, if a neutral side chain (Asp^0 and Glu^0) is required (pH 2.6, net charge of +1). For negatively charged Asp and Glu side chains (Asp$^-$ and Glu$^-$), the scanning of positions 1–5 has been performed on the sequence STVIIT with protected termini (pH 7.4, net charge of −1). Asp and Glu mutants at position 6 have protected termini (pH 7.4, net charge of −1).

[†] Fibrils have been detected only at pH 7.4 (net charge = −1).

[‡] Full β-sheet spectrum. Estimations are lower than 100% because the peptide exhibits a distorted β-sheet signature.

bases with PATTINPROT (17). See Tables 5–13, which are published as supporting information on the PNAS web site.

Representation of Three-Residue Amyloid Motifs (*X*) with Respect to Nonaggregating Sequences (*C*). X is a three-residue motif that matches the amyloid pattern at positions 3, 4, and 5. C is a control sequence that does not have identity with the pattern at positions 3, 4, and 5 and that appears as expected in protein sequences. The probability of a motif $X = ijk$ (P_X) was calculated as the product of the individual amino acid frequencies: $P_X = f_i \times f_j \times f_k$. The individual frequency of an amino acid i (f_i) is the expected frequency for i based on the average composition of 23 proteomes (18). The observed representation of X with respect to C, R_O, was obtained as the ratio between the number of hits found on a database for X (O_X) and C (O_C): $R_O = O_X/O_C$. The expected relative distribution of X with respect to a C (R_E) was calculated as the ratio between the probability of X (P_X) and C (P_C): $R_E = P_X/P_C$. Discordance between expected and found representations was obtained as the ratio between R_E and R_O. If R_E/R_O (± SE) > 1, X is underrepresented respect to the control sequence C; if $R_E/R_O \approx 1$, X appears as expected; and if $R_E/R_O < 1$, X is overrepresented. C is a valid control sequence only if the ratio R_E/R_O, with respect to a different control sequence, is ≈1.

Amyloid Breaker Motif Representation in Proteins. *ijkXX* and *XXijk* are sequence motifs where *ijk* matches the amyloid pattern at positions 3, 4, 5 and X is an amyloid breaker residue [X = Pro (P), Lys (+), Arg (+), Glu (−), and Asp (−)]. *XX* are the following combinations of amyloid breaker residues: ++, − −, +P, P+, −P, and P−. The observed number of hits of an amyloid breaker motif *ijkXX* on a given database is O_{ijkXX} and P_{ijkXX} is its probability. The probability of a motif *ijkYY* (P_{ijkYY}) where $Y \neq$ Arg, Lys, Asp, Glu, or Pro is $P_{ijkYY} = P_{ijk} - P_{ijkXX}$. The ratio between expected probabilities (R) is equal to the ratio between expected values: $R = P_{ijkXX}/P_{ijkYY} = E_{ijkXX}/E_{ijkYY}$. If sequence motifs appeared as expected according to amino acid distribution bases, then the number of *ijk* hits (O_{ijk}) found on a given database should be $O_{ijk} = E_{ijkXX} + E_{ijkYY}$. Because O_{ijk} and R are known, the expected number of hits, E_{ijkXX}, can be calculated as $E_{ijkXX} = O_{ijk}/(R + 1)$. O_{ijk} values are provided in Table 6. To ensure the statistical significance of the differences found between observed (O_{ijkXX}) and expected (E_{ijkXX}) values a paired data Student's *t* test was carried out with the program KALEIDAGRAPH. The df = $n - 1$, with n being the number of observations, i.e., the number of amyloid breaker motifs scanned in this case.

Results and Discussion

Mutation and Amyloid Fibril Formation. The results of the positional scanning mutagenesis on **1** are summarized in Table 1. β-sheet polymerization has been monitored by CD and fibrils detected by EM. Because a good resolution structural model of the fibrils formed by these peptides is still lacking, any discussion about the thermodynamic origin of the effect of mutation on amyloid formation would be difficult and rather speculative. Therefore, we have provided only a descriptive analysis of the sequential dependence found.

Table 1 shows that there is a position dependence on mutation of amyloid fibril formation and that both very tolerant (edges; positions 1, 2, and 6) and restrictive (core; positions 3, 4, and 5) positions can be found within this amyloidogenic sequence. For example, at position 5, only three substitutions give rise to fibril formation. Nevertheless, even among the most tolerant positions, the degree of fibril formation clearly changes on mutation. At position 1, small neutral (Gly, Ala, and Ser) and positively charged amino acids (Lys^+, Arg^+, and His^+) induce the formation of abundant amyloid material. At position 2, polar side chains (Gln, Glu^0, and Thr) provide the most amyloidogenic mutations. At position 6, irrespective of the amyloidogenic substitution made (exceptions are Leu and Glu^-), abundant amyloid material can be detected.

Among the most restrictive positions, 4 is the most permissive regarding the nature of the amino acids that produce a large amount of amyloid deposits (Table 1). Position 3 is much more restrictive in this regard and only some aliphatic amino acids (Val and Leu) cause the formation of numerous fibrils. Position 5 is particularly important for fibril formation because it is restricted to only three hydrophobic amino acids with a high β-sheet propensity (Ile, Phe, and Tyr; ref. 19). The fact that β-branched side chains such as Val and Ile are not permitted at the same place (Val is allowed at position 3, but not at positions 4 and 5, and Ile in the other way around) remarks the high degree of sequence specificity dictated by certain positions. It is also noteworthy that Phe is the only amino acid that allows fibril formation at any position. This result is consistent with the relevance that has been proposed for β-stacking interactions in the self-assembly of amyloid fibrils (16).

The positional dependence found for charged residue substitutions is also very intriguing. Charged side chains are allowed only at the edges of the sequence, but not at the most critical positions. Nevertheless, positive and negative charges are not equally permitted. Whereas negatively charged side chains are allowed at positions 1, 2, and 6, positive charges are compatible with amyloid fibril formation only if they are located at position 1.

Mutation and Onset of β-Sheet Polymerization and Amyloid Fibril Formation. At $t = 0$, the wild-type peptide displays a random coil CD spectrum (Table 1; bold data), that is, most peptide molecules are still in their monomeric conformation (≈80%). At the most tolerant positions (1, 2, and 6), one can find many substitutions that accelerate β-sheet polymerization dramatically (Table 1; italicized annotations). Interestingly, the more restrictive the position is, the less the number of amino acid that are capable of accelerating the process. At position 5, no substitution accelerates the process at all.

Amino acid replacements producing abundant amyloid products are not always the substitutions that allow for a faster β-sheet polymerization (CD, $t = 0$ vs. EM, $t = 1$ month; Table 1). β-sheet estimations show that indeed most mutations that lead to a large amount of fibrils at $t = 1$ month (+++), present a very low percentage of β-sheet population at $t = 0$ (percent $\beta \ll 50\%$), that is, they polymerize slowly (Table 1; underlined annotations). Substitutions at position 6 are the exception: most mutations that accelerate the aggregation process, lead to the formation of abundant amyloid-like material as well. Because all mutants at position 6 contain the pentapeptide STVII, which keeps the capability of forming fibrils (see Fig. 1, which is published as supporting information on the PNAS web site), additional stabilizing interactions at #6 serve to speed up polymerization.

Sequence Pattern for the Detection of Amyloid Fibril-Forming Regions in Proteins. Prediction of amyloid-prone fragments in proteins is a very difficult task because amyloid proteins differ widely in sequence and structure. Nevertheless, investigations carried out during the past years have been found to have some success in extracting some common denominators of amyloid polypeptides (7, 20, 21). These findings encourage the acquisition of sufficient empirical knowledge as a milestone in the development of tools for a reliable prediction.

We have extracted a sequence pattern from these experiments and used it to scan protein sequences for six-residue amyloidogenic stretches. Replacement by charged amino acids at all ionization states has provided us with a different sequence pattern for acidic and neutral pH. The sequence patterns are as follows.

Acidic pH $\{P\}_1$-$\{PKRHW\}_2$-$[VLS(C)WFNQE]_3$-$[ILTYWFNE]_4$-$[FIY]_{5\text{-}}\{PKRH\}_6$.

Neutral pH $\{P\}_1$-$\{PKRHW\}_2$-$[VLS(C)WFNQ]_3$-$[ILTYWFN]_4$-$[FIY]_{5\text{-}}\{PKRH\}_6$.

Residues indicated in square brackets ([]) are those allowed at the position; residues indicated in curly brackets ({}) are those forbidden at the position; an en dash (–) separates each pattern element; and subscripts indicate the sequence position. The use of this sequence pattern to identify six-residue fragments with a tendency to aggregate implies two important assumptions: (*i*) that the relative importance of the sequence positions found in peptide **1** will be general in most six-residue amyloidogenic fragments and (*ii*) that most combinations of the allowed residues at a given position will provide an aggregating motif.

Experimental Validation of the Amyloid Sequence Pattern. To check the generality of the position dependence found in the scanning, we have selected at random some few combinations of the amino acids that lead to a high amyloid feature at the most restrictive positions (core; positions 3, 4, and 5) and combinatorialized the so-called edges: $X_1X_2\mathbf{V}_3\mathbf{I}_4\mathbf{I}_5X_6$, $X_1X_2\mathbf{L}_3\mathbf{N}_4\mathbf{F}_5X_6$, $X_1X_2\mathbf{L}_3\mathbf{E}_4\mathbf{F}_5X_6$, and $X_1X_2\mathbf{W}_3\mathbf{E}_4\mathbf{F}_5X_6$. The X positions are an equimolar mixture of all amino acid substitutions that allow fibrillation at positions 1, 2, and 6. Because every mixture contains >2,000 different sequences, there is a concentration of ≈0.5 μM of each in a 1-mM mixture solution. This concentration is far much lower than the critical concentration of homopolymerization described for some sequence-related peptides (11). This experiment is, therefore, a very stringent test of the aggregating capability of the sequences selected at the core.

At $t = 20$ d, the mixture VII already showed fibril formation, LEF spherical nuclei (fibril precursors) in abundance and LWF and LNF scarce amount of nuclei (see Fig. 2*A*, which is published as supporting information on the PNAS web site). This result suggests the following amyloidogenic ranking of the core sequences tested: VII > LEF > WEF ≈ LNF. To confirm fibril growth from the nuclei observed at $c = 1$ mM, solutions were concentrated ($c \approx 10$ mM) and further incubated for a period of 1 month. EM of the concentrated samples revealed the presence of fibrils of diverse morphology in preparations of VII and LEF and a clear increase of amyloid nuclei in preparations of WEF and LNF (see Fig. 2*B*). Hence, these experiments support the relative prevalence of positions 3, 4, and 5 with respect to positions 1, 2, and 6. Furthermore, they also highlight that highly amyloidogenic amino acid combinations, as short as VII, might drive cofibrillation of a large number of different fragments that just share these three residues.

The second assumption on which a pattern sequence-based detection of aggregating stretches in proteins is based is that any combination of amino acids from the generated pattern produces an aggregating peptide. To reduce the large number of six-residue sequences that can be built from the pattern, we have selected, at each position, those amino acid substitutions that accelerate the polymerization process. Because at position 5, no substitution accelerates aggregation, we have considered both Phe and Ile,

Table 2. Expected vs. observed frequency of amyloidogenic three-residue motifs in the nonredundant protein sequence database

Motif *X*	*P**	$R_{E(\mathrm{ADW})}$[†]	$R_{E(\mathrm{VIS})}$[†]	$R_{O(\mathrm{ADW})}$[‡]	$R_{O(\mathrm{VIS})}$[‡]	$R_E/R_{O(\mathrm{ADW})}$[§] ± 0.178	$R_E/R_{O(\mathrm{VIS})}$[§] ±0.167
VII	2.949e^{-4}	5.765	1.067	2.716	0.526	2.122	2.026
LEF	3.039e^{-4}	5.941	1.099	3.304	0.640	1.798	1.716
WEF	3.377e^{-5}	0.660	0.122	0.524	0.102	1.259	1.202
LNF	1.770e^{-4}	3.458	0.640	4.130	0.801	0.837	0.799
VIV	3.318e^{-4}	6.486	1.200	4.933	0.955	1.315	1.256
ADW	5.116e^{-5}	—	0.186	—	0.194	—	0.960
VIS	2.765e^{-4}	5.405	—	5.159	—	1.045	—

Motif database scanning has been carried out by using the program PATTINPROT (17).

*Motif probabilities have been calculated as the product of the individual amino acid frequencies (see *Methods*).

[†] R_E, the expected distribution of a given motif *X* respect to a control sequence (ADW or VIS): $R_{E(\mathrm{ADW})} = P_X/P_{ADW}$ and $R_{E(\mathrm{VIS})} = P_X/P_{VIS}$.

[‡] R_O, the observed representation of *X* respect to a control sequence (ADW or VIS): $R_{O(\mathrm{ADW})} = O_X/O_{ADW}$ and $R_{O(\mathrm{VIS})} = O_X/O_{VIS}$.

[§] $R_E/R_O > 1$, *X* is underrepresented; $R_E/R_O \approx 1$, *X* appears as expected; $R_E/R_O < 1$, *X* is overrepresented.

residues that are highly amyloidogenic at this position. The resultant pattern for amyloid acceleration is: [VILYNQER]-[AIMF]-[FNQ]-[FN]-[FI]-[LMTND]. From this pattern, we have randomly selected some amino acids ([VIN]-[AI]-[FN]-[FN]-[FI]-[T]) and combined them in such a way that we could compare the aggregating behavior of groups of peptides sharing the same core and presenting different composition at the edges and *vice versa*. The constructed sequences as are follows.

Different core/identical edges:
Group I: NI*FNI*T, NI*NFI*T, NI*NFF*T, NI*FNF*T.
Group II: VA*NFI*T, VA*FNI*T.
Identical core/different edges:
Group III: NI*FNI*T, VA*FNI*T.
Group IV: NI*NFI*T, VA*NFI*T, II*NFI*T.

Samples from all these pattern-derived peptides presented signs of aggregation within hours (see Fig. 3, which is published as supporting information on the PNAS web site); t = 3 h, c = 1 mM, pH 2.6, net charge = +1. Electron micrographs from peptides belonging to groups I and II show that slight changes in the amino acid composition of the amyloid core dramatically affect the degree of amyloidogenecity of peptides that carry the same residues at the edges. The ranking of the constructed amyloidogenic cores in terms of the amount and maturation state of aggregated material detected by EM would be: NFF > FNI > NFI ≥ FNF. Peptides belonging to groups III and IV, which carry the same amyloid cores (FNI and NFI, respectively), showed aggregation products of different morphology as a function of the amino acid composition of the edges. Nevertheless, none of these edge combinations was able to lead to an amount of amyloid material comparable to that found for the most amyloidogenic core (NFF). These results support again that the degree of amyloidogenecity of a given six-residue sequence is mainly determined by the amino acid composition at the core and that the surrounding residues just act as amyloid modulators. This test also indicates that it is likely that most amino acid combinations matching the pattern would generate an aggregating motif, and that aggregating sequences that do not resemble the original peptide **1** at all can be detected or designed by using the pattern. Prediction of the amyloidogenecity degree of pattern-consistent sequences may be, however, a more difficult issue. Amyloid cores constructed by using the most amyloidogenic substitutions have been shown to display a degree of amyloidogenicity remarkably different, ranking from highly to poorly amyloidogenic. Thus, it seems that there is no evident quantitative correlation between individual and combined amino acid propensities to aggregate.

Avoidance of Amyloid Fibril Formation in Proteins: Amyloid Breakers and/or Selection Against Amyloid-Prone Sequences. In the light of the results presented so far, it is tempting to think that protein sequences should have evolved in such a manner that dangerous aggregating motifs, such as VII, are underrepresented respect to innocuous amino acid combinations, and that, in case they must have been conserved for structural/functional reasons, they should be surrounded by other residues that impair their aggregating tendency. These amyloid breakers might be charged amino acids at particular positions of a given amyloidogenic stretch (28), a local excess of charges surrounding the sticky motif (8), and/or Pro, which is a β-sheet breaker (23). To test these hypotheses, we first have scanned some protein databases (17) for the number of VII hits and compared it with the appearance of the other core sequences studied experimentally (LNF, LEF, and WEF). As controls, we have analyzed the frequency of ADW, which does not match the amyloid sequence pattern at all at positions 3, 4, and 5, and VIS, where the amyloid pattern matching is broken at position 5, the most restrictive position within a six-residue amyloid fragment. First of all, Table 2 shows that healthy sequences, that is, controls, appear in protein sequences as expected from protein composition (see ADW vs. VIS and *vice versa*). Second, that VII is indeed underrepresented respect to both controls, and third, that the less amyloidogenic the core sequence is, the more similar is its representation to that expected from the average amino acid composition in proteins. This result implies that whereas the experimental amyloidogenic ranking is VII > LEF > WEF ≈ LNF, their relative representation respect to expected distributions in protein sequences is reverse, LNF > WEF > LEF > VII.

One could wonder whether the low representation of VII in protein sequences is actually general for blocks of three consecutive hydrophobic amino acids. To examine this concern, we have analyzed the frequency of the fully hydrophobic three-residue sequence VIV. VIV mismatches the sequence pattern at only one position (position 5) and the residue that impairs the matching is a Val, a β-branched residue as Ile. Therefore, VIV is a very suitable and demanding motif to probe whether or not the interpretation of the result found for VII is sound. Table 2 shows that, in contrast to VII, VIV is only slightly underrepresented respect to ADW and VIS. This result is indeed consistent with a previous study (24) showing that the frequency of stretches of three consecutive hydrophobic residues in general is the expected in proteins.

To confirm the hypothesis that aggregation driven by highly amyloidogenic three-residue sequences, like VII, might be avoided if they were surrounded by a local excess of charged amino acids and/or Pro, we have also scanned some protein databases for amyloid cores with neighboring amyloid breaker residues. Table 3 shows that at a confidence level higher than 1%, VII is surrounded by charged residues and Pro. Interestingly, the less amyloidogenic the core sequence is, the less frequent amyloid breakers are found at the edges (see Tables 7–13).

Sequence Scanning of Naturally Occurring Amyloid Proteins. As a final test, one might also like to see how well the extracted pattern captures the sequence features of amyloid peptides and proteins

Table 3. Expected vs. observed frequency of the three-residue motif VII surrounded by amyloid breakers in the nonredundant protein sequence database

Motif	O_{VIIXX} (O_{XXVII})*	E_{VIIXX} (E_{XXVII})†
VII + +‡	492	470.07
+ + VII	1,022	470.07
VII − −	1,209	504.24
− − VII	1,739	504.24
P VII	3,471	1648.8
VII P	3,192	1648.8
P + VII	319	184.66
VII + P	244	184.66
P − VII	594	191.26
VII − P	272	191.26
+ P VII	480	184.66
VII P +	300	184.66
− P VII	332	191.26
VII P −	336	191.26
Mean difference§	518.006	
t value§	3.2572	
t probability§	0.006242	

Motif database scanning has been carried out by using PATTINPROT (17).

*O_{VIIXX} (O_{XXVII}), the observed number of hits of an amyloid breaker motif on the database.

†E_{VIIXX} (E_{XXVII}), the expected number of hits of an amyloid breaker motif based on the average composition of proteins (see *Methods*) and the total number of VII hits found on the database (O_{VII} = 37,473).

‡VII + +, the only amyloid breaker motif found as expected from amino acid composition.

§Student's *t* test.

that form fibrillar aggregates *in vitro* and/or *in vivo*. Since for some of these proteins the regions with aggregating capability have been already identified experimentally, the sequence scanning of these proteins constitutes a very convenient way to validate the detection of amyloidogenic regions in protein sequences based on this pattern (Table 4).

First, we have scanned sequences of proteins that do not form amyloid fibrils *in vivo*, but that have been shown to do so under denaturing conditions (acidic pH and/or cosolvents). Dobson and coworkers (25) have shown that the aggregation rate of the muscle acylphosphatase is very sensitive to mutation at the region 87–98. This region contains, in fact, a six-residue fragment that matches the pattern. A peptide comprising the last two β-strands of the bacterial cold shock protein also fibrillates (26). Based on our pattern prediction, the first of these two β-strands (residues 46–51) would be responsible for the aggregation of this peptide. Although the SH3 domain of phosphatidylinositol-3′-kinase (PI3-SH3) and the α-spectrin SH3 domain share a high degree of sequence and structural homology, they present a very different aggregating behavior: whereas PI3-SH3 forms amyloid-like fibrils under acidic conditions, the α-spectrin SH3 does not (27). Consistent with these results, we have found that for the latter no fragment has identity with the pattern and that, in contrast, PI3-SH3 contains two fragments that match the pattern only at acidic pH.

The N-terminal region (residues 90–120) of the cellular prion protein ($\mathrm{PrP}^{\mathrm{C}}$) is disordered and protected from protease digestion in the oligomeric scrapie isoform ($\mathrm{PrP}^{\mathrm{Sc}}$), which has given support to the notion that a structural change in this region is the minimal requirement for the transition $\mathrm{PrP}^{\mathrm{C}} \rightarrow \mathrm{PrP}^{\mathrm{Sc}}$ (28). However, this fragment alone is not sufficient to explain the large secondary β-structure content of $\mathrm{PrP}^{\mathrm{Sc}}$ (29), indicating that conformational rearrangements in the ordered C-terminal region of $\mathrm{PrP}^{\mathrm{C}}$ are required to produce the priogenic intermediate (30). In fact, most amyloidogenic stretches detected with the pattern are found within this region (Table 4). Two of these fragments, covering residues 171–176 and 178–183, are located in helix II, in concordance with

Table 4. Predicted vs. experimental amyloid-forming peptides in amyloid proteins

Protein*	Predicted (acidic pH)	Predicted (neutral pH)	Experimental region and conditions
Muscle acetylphosphatase (Acp)	^{90}EYSNFS^{95}	^{90}EYSNFS^{95}	Residues 87–98; pH 5.5, 25% vol/vol TFE (25)
PI3-SH3	^{72}TYVEYI^{77}	No identity†	pH 2.0 (5, 27)
	^{73}YVEYIG^{78}		
α-spectrin SH3	No identity†	No identity†	—‡ (27)
Bacterial cold shock protein (CspB)	^{46}VSFEIV^{51}	No identity†	Residues 36–67; pH 4.0, 50% AcN (26)
PrP Human	^{8}MLVLFV^{13}	^{8}MLVLFV^{13}	Helix II (21, 31)
	^{171}NQNNFV^{176}	^{171}NQNNFV^{176}	C-terminal region (30)
	^{178}DCVNIT^{183}	^{178}DCVNIT^{183}	
	^{194}KGENFT^{199}	—	
	^{231}SMVLFS^{236}	^{231}SMVLFS^{236}	
	^{240}VILLIS^{245}	^{240}VILLIS^{245}	
	^{244}ISFLIF^{249}	^{244}ISFLIF^{249}	
	^{245}SFLIFL^{250}	^{245}SFLIFL^{250}	
	^{247}LIFLIV^{252}	^{247}LIFLIV^{252}	
Sup35 (yeast prion)	^{9}NQQNYQ^{14}	^{9}NQQNYQ^{14}	^{7}GN**NQQNY**^{13} (32, 33)
	^{45}YYQNYQ^{50}	^{45}YYQNYQ^{50}	
	^{102}KNFNYN^{107}	^{102}KNFNYN^{107}	
IAPP rat	No identity†	No identity†	—‡
IAPP human	^{19}SSNNFG^{24}	^{19}SSNNFG^{24}	^{20}**SNNFGAIL**^{27} (16)
Aβ (1–42) peptide	^{16}KLVFFA^{21}	^{16}KLVFFA^{21}	^{14}HQ**KLVFFAED**^{23} (37, 38)
ApoA	^{8}LAVLFL^{13}	^{8}LAVLFL^{13}	N-terminal region (residues 1–93) (34)
	^{77}VTSTFS^{82}	^{77}VTSTFS^{82}	
	^{91}VTQEFW^{96}		
	^{256}ALEEYT^{261}		
Human β-2-microglobulin (β_2m)	^{63}FYLLYY^{68}	^{63}FYLLYY^{68}	Residues 61–70 (35, 36)
	^{64}YLLYYT^{69}	^{64}YLLYYT^{69}	

Protein sequences have been scanned for six-residue fragments matching the amyloid pattern (see *Methods*).

*The fragments identified with the pattern for some proteins, such as α-synuclein, τ, lysozyme, etc., have not yet been shown to form amyloid fibrils experimentally. These proteins have not been included in this table.

†No fragment within the protein sequence matches the pattern.

‡—, The protein does not aggregate under any experimental conditions tested.

the amyloidogenic behavior shown for peptides comprising this helix in comparison to helices I and III, which do not have any fragment matching the pattern (31).

Gln/Asn repeating sequences interspersed with other amino acids have been suggested to promote and stabilize prion formation in the yeast protein Sup35 (32). Such sequences, i.e., a fragment of the N-terminal domain of Sup35 (33), can be also detected with the pattern (Table 4).

Scanning of the human islet amyloid polypeptide (IAPP) detects a stretch that is within a IAPP fragment that is amyloidogenic and cytotoxic (14). The inhibitory effect of the mutation Phe23Ala on the amyloidogenicity of IAPP (16) can be also predicted with the pattern, because replacement of Phe-23 by Ala impairs matching with the pattern at position 5. Consistent with the fact that rat IAPP does not form amyloid fibrils, no fragment of its sequence matches the pattern (16).

For apolipoprotein A (ApoA), involved in hereditary systemic amyloidosis, three amyloid stretches are predicted within the region corresponding to residues 8–96 (Table 4). In fact, most mutations described in ApoA are within the N-terminal portion of the protein (1–93), which represents the proteolysis fragment that is incorporated into amyloid deposits (34).

Data from Radford and coworkers (35) indicate that the amyloidogenic intermediate of the human β_2-microglobulin (β_2m) retains a stable domain involving the five β-strands of the native fold, with strand E (residues 61–70) being the most stable. Our scanning predicts that strand E is indeed the only amyloidogenic fragment of β_2m. Consistent with this, out of all synthesized molecules comprising β_2m strands, only peptides containing the sequence of strand E form amyloids (36).

The only fragment of Aβ (1–42) that agrees with the pattern consists of residues 16–21. Interestingly, it has been shown that residues 16–20 in Aβ are essential for Aβ polymerization (37). Furthermore, Hecht and coworkers (38) have found that Aβ solubility is very sensitive to mutations within this region. In particular, mutation Phe19Ser, which impairs identity with the pattern at position 4, gives rise to the highest solubility of the Aβ (1–42) peptide *in vivo* (38).

Conclusions

The large amount of protein sequences provided by large-scale genomics and proteomics initiatives is demanding for computational methods to predict not only structure and function but also disease. A reliable identification of amyloid motifs in proteins should have a great impact in the development of antiamyloid therapeutics. Putative amyloid fragments could be synthesized as small peptides and be assayed against large libraries of drugs and/or be used as scaffold for the rational design of molecules capable of interfering with polymerization from the region(s) identified as amyloidogenic.

The sequence pattern described here represents a tool for developing an algorithm for the prediction of amyloidogenic fragments in proteins. Because the sequence space explored in this work is very small, we are conscious that many dangerous motifs cannot be detected unless a more complete pattern is provided. Therefore, similar experiments on unrelated amyloid sequences would be very useful to refine this amyloid fingerprint. Nevertheless, this pattern has been shown to contain enough relevant information to identify amyloid fragments from different sources, allowing for the detection and design of sequences that are very different from the scanned molecule. These results suggest that this pattern in combination with other experimental observations may be useful for identifying new proteins prone to forming disease-causing aggregates.

This work is supported by a biotechnology grant from the European Union (to L.S.) and by a European Union Marie Curie Fellowship (to M.L.d.l.P.).

1. Dobson, C. M. (1999) *Trends Biochem. Sci.* **24,** 329–332.
2. Rochet, J. C. & Lansbury, P. T., Jr. (2000) *Curr. Opin. Struct. Biol.* **10,** 60–68.
3. Thirumalai, D., Klimov, D. K. & Dima, R. I. (2003) *Curr. Opin. Struct. Biol.* **13,** 146–159.
4. Sipe, J. D. & Cohen, A. S. (2000) *J. Struct. Biol.* **130,** 88–98.
5. Guijarro, J. I., Sunde, M., Jones, J. A., Campbell, I. D. & Dobson, C. M. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 4224–4228.
6. Chiti, F., Webster, P., Taddei, N., Clark, A., Stefani, M., Ramponi, G. & Dobson, C. M. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 3590–3594.
7. West, M. W., Wang, W., Patterson, J., Mancias, J. D., Beasley, J. R. & Hecht, M. H. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 11211–11216.
8. López de la Paz, M., Goldie, K., Zurdo, J., Lacroix, E., Dobson, C. M., Hoenger, A. & Serrano, L. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 16052–16057.
9. Villegas, V., Zurdo, J., Filimonov, V. V., Aviles, F. X., Dobson, C. M. & Serrano, L. (2000) *Protein Sci.* **9,** 1700–1708.
10. Hammarstrom, P., Jiang, X., Hurshman, A. R., Powers, E. T. & Kelly, J. W. (2002) *Proc. Natl. Acad. Sci. USA* **99,** Suppl. 4, 16427–16432.
11. Liemann, S. & Glockshuber, R. (1999) *Biochemistry* **38,** 3258–3267.
12. Swietnicki, W., Petersen, R. B., Gambetti, P. & Surewicz, W. K. (1998) *J. Biol. Chem.* **273,** 31048–31052.
13. Uversky, V. N., Gillespie, J. R. & Fink, A. L. (2000) *Proteins* **41,** 415–427.
14. Tenidis, K., Waldner, M., Bernhagen, J., Fischle, W., Bergmann, M., Weber, M., Merkle, M. L., Voelter, W., Brunner, H. & Kapurniotu, A. (2000) *J. Mol. Biol.* **295,** 1055–1071.
15. von Bergen, M., Friedhoff, P., Biernat, J., Heberle, J., Mandelkow, E. M. & Mandelkow, E. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 5129–5134.
16. Azriel, R. & Gazit, E. (2001) *J. Biol. Chem.* **276,** 34156–34161.
17. Combet, C., Blanchet, C., Geourjon, C. & Deleage, G. (2000) *Trends Biochem. Sci.* **25,** 147–150.
18. Berezovsky, I. N., Kirzhner, A., Kirzhner, V. M., Rosenfeld, V. R. & Trifonov, E. N. (2003) *J. Biomol. Struct. Dyn.* **21,** 317–326.
19. Minor, D. L., Jr., & Kim, P. S. (1994) *Nature* **367,** 660–663.
20. Turner, W. G. & Finch, J. T. (1992) *J. Mol. Biol.* **277,** 1205–1223.
21. Kallberg, Y., Gustafsson, M., Persson, B., Thyberg, J. & Johansson, J. (2001) *J. Biol. Chem.* **276,** 12945–12950.
22. Otzen, D. E., Kristensen, O. & Oliveberg, M. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 9907–9912.
23. Adessi, C. & Soto, C. (2002) *Drug Dev. Res.* **56,** 184–193.
24. Schwartz, R., Istrail, S. & King, J. (2001) *Protein Sci.* **10,** 1023–1031.
25. Chiti, F., Taddei, N., Baroni, F., Capanni, C., Stefani, M., Ramponi, G. & Dobson, C. M. (2002) *Nat. Struct. Biol.* **9,** 137–143.
26. Gross, M., Wilkins, D. K., Pitkeathly, M. C., Chung, E. W., Higham, C., Clark, A. & Dobson, C. M. (1999) *Protein Sci.* **8,** 1350–1357.
27. Ventura, S., Lacroix, E. & Serrano, L. (2002) *J. Mol. Biol.* **322,** 1147–1158.
28. Peretz, D., Williamson, R. A., Matsunaga, Y., Serban, H., Pinilla, C., Bastidas, R. B., Rozenshteyn, R., James, T. L., Houghten, R. A., Cohen, F. E., *et al.* (1997) *J. Mol. Biol.* **273,** 614–622.
29. Wille, H., Zhang, G. F., Baldwin, M. A., Cohen, F. E. & Prusiner, S. B. (1996) *J. Mol. Biol.* **259,** 608–621.
30. Hornemann, S. & Glockshuber, R. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 6010–6014.
31. Thompson, A., White, A. R., McLean, C., Masters, C. L., Cappai, R. & Barrow, C. J. (2000) *J. Neurosci. Res.* **62,** 293–301.
32. Michelitsch, M. D. & Weissman, J. S. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 11910–11915.
33. Balbirnie, M., Grothe, R. & Eisenberg, D. S. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 2375–2380.
34. Hamidi Asl, K., Liepnieks, J. J., Nakamura, M., Parker, F. & Benson, M. D. (1999) *Biochem. Biophys. Res. Commun.* **257,** 584–588.
35. McParland, V. J., Kalverda, A. P., Homans, S. W. & Radford, S. E. (2002) *Nat. Struct. Biol* **9,** 326–331.
36. Jones, S., Manning, J., Kad, N. M. & Radford, S. E. (2003) *J. Mol. Biol.* **325,** 249–257.
37. Tjernberg, L. O., Callaway, D. J., Tjernberg, A., Hahne, S., Lilliehook, C., Terenius, L., Thyberg, J. & Nordstedt, C. (1999) *J. Biol. Chem.* **274,** 12619–12625.
38. Wurth, C., Guimard, N. K. & Hecht, M. H. (2002) *J. Mol. Biol.* **319,** 1279–1290.
39. Jiménez, M. A., Muñoz, V., Rico, M. & Serrano, L. (1994) *J. Mol. Biol.* **242,** 487–496.
40. Greenfield, N. & Fasman, G. D. (1969) *Biochemistry* **8,** 4108–4116.