Identifying the amylome, proteins capable of forming amyloid-like fibrils

Lukasz Goldschmidt[a], Poh K. Teng[a], Roland Riek[b], and David Eisenberg[a,1]

[a]Howard Hughes Medical Institute, University of California Los Angeles–Department of Energy Institute for Genomics and Proteomics, Los Angeles, CA 90095-1570; and [b]Laboratory of Physical Chemistry, Eidgenössiche Technische Hochschule Zurich, CH-8093 Zürich, Switzerland

Contributed by David S Eisenberg, January 4, 2010 (sent for review December 3, 2009)

The amylome is the universe of proteins that are capable of forming amyloid-like fibrils. Here we investigate the factors that enable a protein to belong to the amylome. A major factor is the presence in the protein of a segment that can form a tightly complementary interface with an identical segment, which permits the formation of a steric zipper—two self-complementary beta sheets that form the spine of an amyloid fibril. Another factor is sufficient conformational freedom of the self-complementary segment to interact with other molecules. Using RNase A as a model system, we validate our fibrillogenic predictions by the 3D profile method based on the crystal structure of NNQQNY and demonstrate that a specific residue order is required for fiber formation. Our genome-wide analysis revealed that self-complementary segments are found in almost all proteins, yet not all proteins form amyloids. The implication is that chaperoning effects have evolved to constrain self-complementary segments from interaction with each other.

3D profile | ribonuclease A | Rosetta energy | steric zipper

Seventy-five years ago, the pioneering biophysicist William Astbury speculated that every protein might have a fibrous state as well as a globular state (1). Astbury was the first to describe the cross-beta fibril diffraction pattern, now accepted as the definitive signature of the amyloid state of proteins. Astbury's observation was on a denatured protein, albumin in poached egg white. Today it is established that amyloid diseases, including Alzheimer's and prion diseases, are associated with elongated, unbranched protein fibrils (2, 3). However, functional proteins are also found in the amyloid state. These include the egg stalk of the green lace-wing fly (4), the Pmel17 protein associated with skin pigmentation (5), and a large number of secretory hormones (6). Conversely, in the past decade, Pertinhez et al. (7) and others (8–10) have shown that many globular proteins can be converted to the amyloid state by a variety of denaturing processes, suggesting that conversion may be generally applicable to all proteins. So the question arises, to what extent is this conjecture true? That is, how large is the amylome?

Computer algorithms have been proposed to answer a somewhat broader question: What is the aggregation propensity of a given protein sequence? Aggregates, in general, include amyloid-like fibrils but also other types of fibrils and nonfibrillar aggregates. TANGO (11) identifies beta-aggregating regions of proteins by using a statistical mechanics algorithm based on the physico-chemical principles of beta-sheet formation. For each residue, it calculates the energy of structural states derived from statistical and empirical considerations and then computes the occupancy of the beta-aggregation conformational state. Although beta-aggregation propensity by itself is not necessarily indicative of amyloid formation, it plays a major role in determining the tendency to ultimately form organized structures such as amyloid fibrils. However, as stated by the authors, TANGO cannot be used to quantitatively compare aggregation propensities between different proteins and is unable to accurately predict low levels of aggregation propensity. The Dobson and Vendruscolo groups developed an algorithm that predicts the regions of protein sequences that are most important in promoting aggregation and amyloid formation. Zyggregator (12), an implementation of this algorithm, computes a Z_{agg} score profile, taking into account both the intrinsic aggregation propensity computed from the protein sequence and the structural protection provided by the folded form of the protein. PASTA, the prediction of amyloid structure aggregation algorithm (13), uses a pairwise energy function that computes the propensity of two residues found within a beta sheet facing one another on neighboring strands. It is based on the key assumption that a universal mechanism is responsible for beta-sheet formation in globular proteins and fibrillar aggregates. PASTA's ability to predict the registry of the intermolecular hydrogen bonds formed between amyloidogenic segments allows it to identify the portions of the sequence forming the cross-beta core as well as to discriminate whether the intermolecular beta-strands are parallel or antiparallel. These three methods are sequence-based and do not take 3D structural features directly into account.

Here we focus on the factors that permit a protein to convert to the amyloid state. We extend the 3D profile method (14) to computationally identify, within all putative proteins of three genomes, segments with high fibrillation propensity (HP) that can form a "steric zipper"—two self-complementary beta sheets, giving rise to the spine of an amyloid fibril. Our approach differs from those discussed above in that it relies mainly on structural information to evaluate the likelihood that a particular sequence can form fibrils, in contrast to previous algorithms that rely mainly on sequence information. For structures deposited in the Protein Data Bank (PDB), we also establish the localization and geometry of such segments in folded proteins. Using bovine pancreatic ribonuclease A (RNase A) as a model system, we experimentally validate the accuracy of our predictions and investigate the effect of sequence and residue composition. We demonstrate that our prediction method of self-complementary, fibrillizing segments within proteins is effective; that fibrillation propensity depends on the specific residue sequence; and that self-complementary segments on the surface of proteins are rare, suggesting that chaperoning effects have evolved to constrain such segments from interaction with each other (15).

Results

Identification of Protein Segments Having High Propensity for Fibrillation. In previous studies, we found that particular short (4–10 residue) segments of proteins are capable of forming amyloid-like fibrils (16–19). Atomic structures of fibril-like crystals formed by these segments revealed that these segments are self-complementary, stacking into pairs of beta sheets, whose side chains extend and interdigitate. On the basis of the first two atomic structures (of segments GNNQQNY and NNQQNY), we

Author contributions: L.G., P.K.T., R.R., and D.E. designed research; L.G. and P.K.T. performed research; L.G., P.K.T., R.R., and D.E. analyzed data; and L.G. and D.E. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: david@mbi.ucla.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0915166107/DCSupplemental.

developed a structure-based, computational method for identification of other short segments that form fibrils (14). In this 3D profile method, the sequences of putative amyloid-forming proteins are scanned by threading them on the backbone of the known crystal structure of the segment NNQQNY. Segments that can form a similar, self-complementary steric zipper structure with a low energy, as assessed by the RosettaDesign potential energy function (20), are judged as being capable of fibril formation.

On the basis of a set of 16 published hexapeptide zipper crystal structures determined in our group (*SI Appendix*), we have established an energetic threshold for HP. We predict segments to be self-complementary and have HP when their Rosetta energy is at or below −23 kcal/mol after their sequence is threaded on the NNQQNY backbone by using the 3D profile method. Of the 16 segments, 13 have Rosetta energies below −23 kcal/mol. No segments have energies above our permissive threshold of −19 kcal/mol. TANGO classified only three of these segments, VQIVYK, GGVVIA, and MVGGVV, as having aggregation propensity.

Experimental Assessment of the 3D Profile Method for Identifying Segments with High Propensity for Fibrillation. To validate our algorithm, we examined the fibrillation of RNase A segments predicted to form fibrils and others predicted not to form fibrils (Fig. 1). We selected six hexapeptides that were predicted to fibrillize because in the putative fibrillar state their energies fall below −23 kcal/mol, and we confirmed by electron microscopy that they indeed fibrillize. Additionally, we tested two other segments with predicted energies considerably above −23 kcal/mol and verified that they do not fibrillize in any tested condition (Fig. 1 and *SI Appendix*). Next we examined three shuffled hexapeptide pairs, where we predicted only one member of each pair to fibrillize. Both members of a pair share identical amino acid composition and differ only in the order of the residues within the segments. We monitored fibrillation by electron microscopy and confirmed that after a 4–5-week incubation only the predicted HP segments fibrillized, whereas their shuffled, higher-energy mates did not (Fig. 1 and *SI Appendix*). These experiments suggest that (*i*) the specific residue sequence within the segment determines fibrillation, rather than amino acid composition alone, and (*ii*) a value of −23 kcal/mol for the threshold of prediction of fibrillation is effective for the classification of segment propensities.

Genome-Wide Prediction of Segments with High Propensity for Fibrillation. We predicted fibrillation propensities within all annotated ORFs in the *Escherichia coli*, *Saccharomyces cerevisiae*, and *Homo sapiens* genomes (4,094, 5,814, and 38,927 sequences, respectively; Table 1) as well as within all proteins in a nonredundant set of structures in the PDB on the basis of 50% sequence identity (12,836 sequences). To obtain an initial estimate of the HP segment enrichment, we analyzed all four datasets by using the simplified triplet method (which takes into account only the interactions of the three inward-facing residues at the steric zipper interface; see *Methods*), because the size of the two larger genomes makes the use of the more accurate and slower 3D profile method infeasible. Our benchmarks of the triplet method indicate that it introduces an average error of 4 kcal/mol per hexapeptide and has a tendency to overestimate propensity. Therefore this method is mostly useful for the comparison of the HP enrichment levels between datasets and provides only an approximate estimate of the exact enrichment percentage. We find that the enrichment for high propensity segments (the percentage of segments that are HP) computed by using the triplet method is very similar for all four datasets (18.8–20.0%).

We next analyzed the *E. coli* genome and the nonredundant set of the PDB with the more accurate 3D profile method and determined the true HP segment enrichment in these two datasets to be 15.1% and 14.0%, respectively. On the basis of the initial estimates by the triplet method, we expect the enrichment in the *S. cerevisiae* and *H. sapiens* genomes to be similar. Nearly all ORFs contain at least one HP segment, and on average 14–15% of all segments within any given protein have HP. Strikingly, proteins with known structure also contain one and often several HP segments, yet almost none of these proteins readily form fibrils under physiological conditions. Clearly there must be mechanisms in place to protect these proteins from fibrillation.

Localization of HP Segments. Having discovered that HP segments are common, we asked where these segments are in primary sequence and 3D space in proteins of known structure. We did not detect a location preference of HP segments in the primary sequence and only a slight preference for some secondary structure elements. Rather, HP segments tend to be distributed throughout protein sequences, with no strong tendency to cluster at the termini, or away from them. Often, however, high propensity segments are directly flanked by charged residues such as aspartic acid and arginine (11% and 14% of the time by Asp and Arg, respectively), similar to the findings by Rousseau et al. (21).

Further, we analyzed all segments that could be unambiguously assigned to a secondary structure class by the Definition of Secondary Structure of Proteins (DSSP) (at least five residues have the same class, α-helix, beta strand, or coil) for their HP enrichment. Not surprisingly, beta strands have the highest HP segment enrichment (28.2%) followed by α-helices (18.8%). Coil regions, commonly found in flexible surface-exposed regions, contain the fewest HP segments (8.8% enrichment), which suggests that proteins have evolved to display fewer HP segments in flexible, surface-exposed regions.

Solvent Exposure and Conformation of HP Segments in Proteins of Known Structure. The suggestive result of the preceding paragraph—that proteins have evolved to contain relatively few HP segments in surface-exposed segments—can be tested, as follows. To establish the location of HP segments in 3D structures, we analyzed proteins of known structure in a nonredundant set of the PDB. The majority of segments with HP are buried and therefore inaccessible for intermolecular contacts (Table 2). We find that only 5.3% of high propensity, analyzable segments are surface-exposed and that effectively none (<0.10%) of these exposed HP segments have beta-strand compatible geometry. Furthermore, disordered regions of proteins, such as those that could not be resolved by x-ray crystallography, contain one-third fewer HP segments (9.6%; *SI Appendix*) than the entire protein. We conclude that, in folded proteins, segments with HP are nearly always protected from interaction with identical segments in other molecules.

Function of HP Segments. Because HP segments appear to be so widespread, we hypothesize that there is little evolutionary pressure to remove or alter them or that these segments cannot be removed because of functional constraints. To attempt to address the latter hypothesis, we examined all HP segments for the presence of functional residues, such as highly conserved residues or annotated active site residues in UniProt. Our analysis did not detect a significant difference in the enrichment of conserved or active site residues within HP segments compared to non-HP segments. That is, the difference we observe is less than the overall difference in enrichment between the various datasets reported in Table 1. However, we note that typically only a single residue within a segment has an active site annotation or shows significant conservation. This signal may be insufficient for our current methodology to correlate the entire segment with function.

Amino Acid Composition of HP Segments. We asked whether or not the localization and propensity of HP segments are solely due to their amino acid composition. HP segments are enriched for

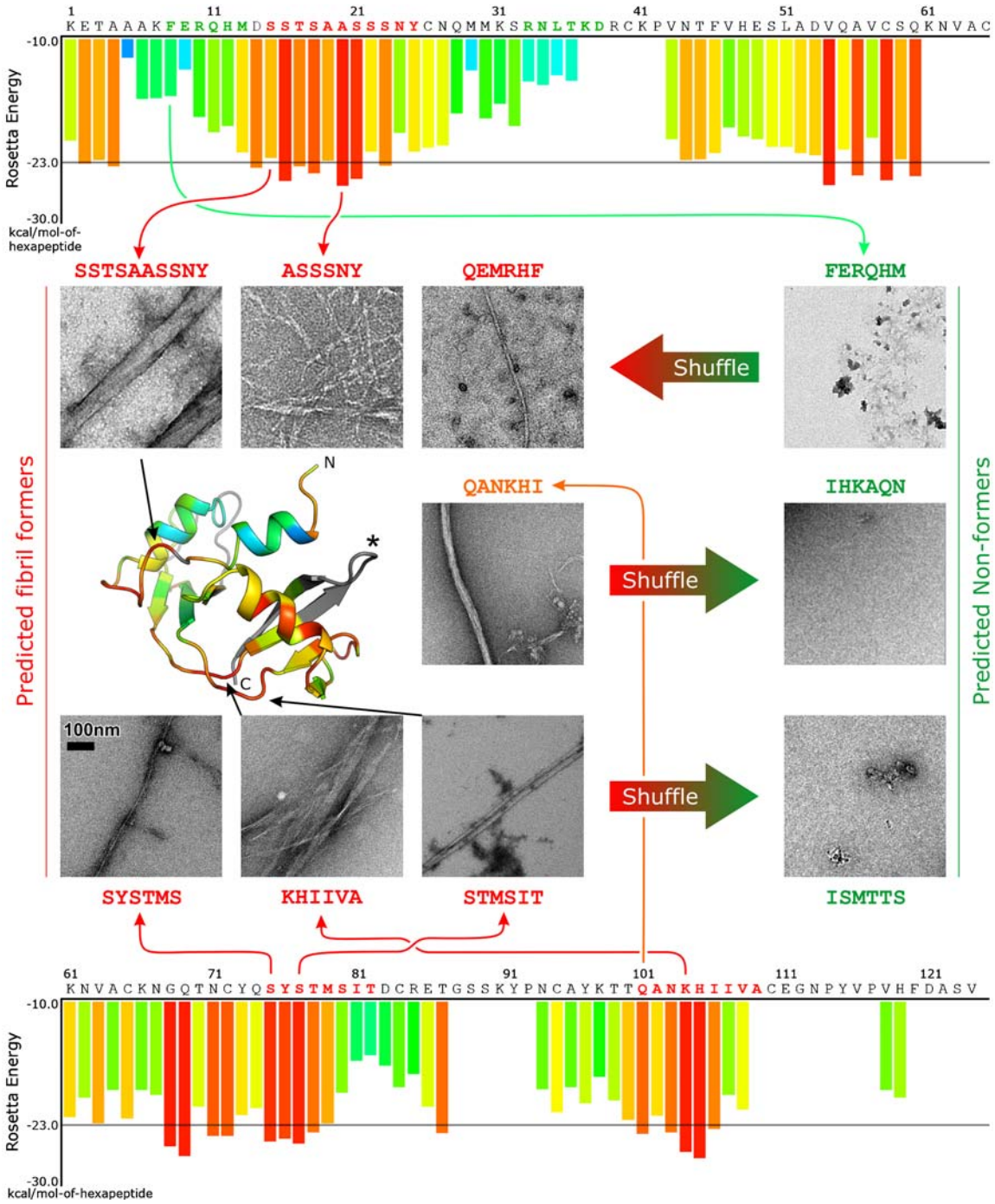

Fig. 1. Validation of the 3D profile method for prediction of fibrillizing segments. The top and bottom panels show the predicted energy for fibrillation of every six-residue segment of RNase A. Red histogram bars represent hexapeptides with energy below −23 kcal/mol and are predicted to form fibrils. Seven such segments are shown, and each forms fibrils, as shown by the electron micrographs in the central panels. The black arrows indicate where three of these segments lie in the 3D structure of RNase A. This cartoon structure, like the histogram of energies, is colored with warmer colors representing a greater propensity for fibrillation. Blue and green segments are of higher energy and are predicted not to form fibrils. FERQHM is one of several segments predicted and experimentally confirmed not to form fibrils (first row of micrographs, right). However, when the residues of this segment were rearranged to QEMRHF (Table S2), the energy of the rearranged segment falls below −23 kcal/mol; QEMRHF is thus predicted to form fibrils, and it does (first row of micrographs). In addition, the fibril-forming segments QANKHI and STMSIT were rearranged to IHKAQN and ISMTTS, respectively, two sequences predicted and confirmed not to form fibrils (second and third row of micrographs). Notice that the longer SSTSAASSNY segment contains the SSTSAA segment, which forms fibrils and is capable of forming a steric zipper (Sawaya 2007). Taken together, the fibrillizing behavior of these 10 segments suggests that a threshold of −23 kcal/mol is appropriate for predicting fibrillation. The * indicates the C-terminal hinge loop, discussed in the text.

valine, isoleucine, alanine, and serine, whereas charged residues such as aspartate, glutamate, and arginine are less favored, particularly at the zipper interface. In our computational analysis, charged residues are disfavored at the core of the steric zipper because of unfavorable solvation energy when such residues are buried. We expect that, in solution, an arrangement of stacked charges in a fibril is also disfavored, although our current computational setup does not take this effect directly into account.

Table 1. Abundance of HP segments in ORFs of three genomes and in a nonredundant set of the PDB on the basis of 50% sequence identity. HP segments are common in ORFs in genomes. All four datasets contain a comparable percentage of HP segments, and almost all ORFs contain at least one such segment. Fibrillogenic propensities were computed by using the simplified but less accurate triplet method (see *Methods*) and for the smaller *E. coli* and PDB50 datasets by using the 3D profile method.

Genome	ORFs	Segments	HP Segments	% HP segments	HP ORFs	% HP ORFs
By triplet method						
E. coli	4,094	1,286,485	282,372	22.0	4,094	100.0
S. cerevisiae	5,814	2,864,118	622,601	21.7	5,813	99.98
H. sapiens	38,927	19,737,602	3,712,536	18.8	38,913	99.96
PDB	26,617	6,037,431	1,205,513	20.0	26,534	99.69
By 3D profile method						
E. coli	4,094	1,286,485	194,042	15.1	4,090	99.90
PDB50	12,836	2,847,049	397,939	14.0	12,672	98.72

These considerations suggest that HP segments will be more likely buried than exposed; however, the hydrophobicity bias alone does not explain the rarity of surface-exposed HP segments in proteins.

To verify that residue composition alone does not account for segment propensity or exposure, we shuffled the sequences *in silico* of all exposed and nonexposed HP segments, as well as exposed and nonexposed non-HP segments in proteins of known structure (crystallographically determined structures in the PDB). In this experiment, if a shuffled sequence of a given segment was present in any other structure, that structure was used to compute the exposure of the shuffled segment. Fig. 2*A* shows the distribution of fibrillation propensity (Rosetta energy) and the exposure of all unshuffled segments in known structures. We observe a small trend for HP segments to be preferentially buried. An equivalent propensity versus exposure distribution heat map is shown in Fig. 2*B* for segments shuffled from originally buried, high propensity segments. The fibrillation propensity of many of these segments (44%) is disrupted by shuffling, as shown by their energies increasing to above our threshold of −23 kcal/mol. Some segments (30%) also become exposed, and about half of these exposed segments (16% of total) lose their HP.

Combined, these experiments demonstrate that HP requires a specific residue order and that a compatible amino acid composition is necessary but not sufficient. Our experimental results on a few shuffled segment pairs (Fig. 1) further emphasize the importance of residue order. It is noteworthy that a single residue change can greatly affect propensity. For example, the F19R mutation in A-beta that changes the sequence from QKLVFF to QKLVRF raises the energy from −24.0 to −18.4 kcal/mol. Conversely, the F19A mutation of the same residue lowers the energy to −25.2 kcal/mol.

Table 2. Enrichment, exposure, and geometry of high propensity segments in proteins with known structure (nonredundant set of the PDB on the basis of 50% identity). On the basis of our definition, about half of all segments are surface-exposed; 15% of all segments have HP. There are, however, few surface-exposed HP segments (5.3%) and almost no segments (<0.10%) that also have a beta-strand compatible geometry that is necessary for the formation of intermolecular, self-complementary beta sheets. Percentages are based on the total analyzable structures—those that were successfully analyzed with DSSP (about 80% of the entire dataset).

	Segments	% segments	Proteins	% proteins
PDB50	2,848,309		12,846	
Analyzable structures	2,475,775	100	10,269	100
Proline-free segments	1,965,517	79.4	10,267	100
Exposed	1,369,993	55.3	10,269	100
High propensity	365,401	14.8	10,195	99.3
β-strand compatible geometry	450,637	18.2	8,170	79.6
HP + exposed	131,509	5.3	10,048	97.8
HP + exposed + geometry	2157	<0.10	1193	11

Discussion

Reliability of the 3D Profile Method for Identifying Fibrillizing Segments. Our experiments with segments of RNase A, summarized in Fig. 1, suggest that the 3D profile algorithm is effective in identifying fibrillizing segments of proteins. In previous work, the method had predicted several dozen segments from about 10 proteins that formed fibrils or fibril-like microcrystals (14, 19, 22, 23). The RNase A studies go further in helping to define a threshold for prediction of fibrillation, as well as demonstrating that higher-energy segments do not fibrillize, even at concentrations of 10 mM.

Sequence Versus composition. Our shuffling experiments suggest that the tendency to form amyloid-like fibrils is strongly sequence-dependent and relatively insensitive to amino acid composition. That is, when the sequence of a fibrillizing segment is shuffled, the rearranged sequence loses its tendency to form fibrils. Conversely, when the sequence of a nonfibrillizing segment is rearranged to one with a low energy, it converts to a fibril former.

The Segment Amylome. We have estimated the number of different HP protein segments capable of forming amyloid-like fibrils, which we term the segment amylome (Table 1). ORFs in all three examined genomes and in protein structures deposited in the PDB contain at least one and often several analyzable segments capable of forming fibrils. The enrichment for such HP segments is comparable in all four datasets (14–15%). Considering the limited number of proteins found so far to form fibrils at physiological conditions and especially considering that most proteins in the PDB do not readily fibrillize, this enrichment is rather high. Yet, the ever-growing body of evidence gathered in our group clearly demonstrates that HP segments identified by the 3D profile method do indeed form fibrils. Undeniably, the context of these HP segments within the protein is of key importance in controlling their fibrillation propensity.

The Amylome. The propensity of a protein to form fibrils depends in part on the presence of HP segments within its sequence, but there are clearly other influential factors. To uncover these factors, we studied the placement and conformation of HP segments within proteins of known structure. As reported in our results, we found that 95% of HP segments are buried, and, of those exposed to solvent, >99.9% are twisted and thus incompatible with the very straight geometry of the NNQQNY backbone (Table 2). Within all proteins of known structure, there are virtually no

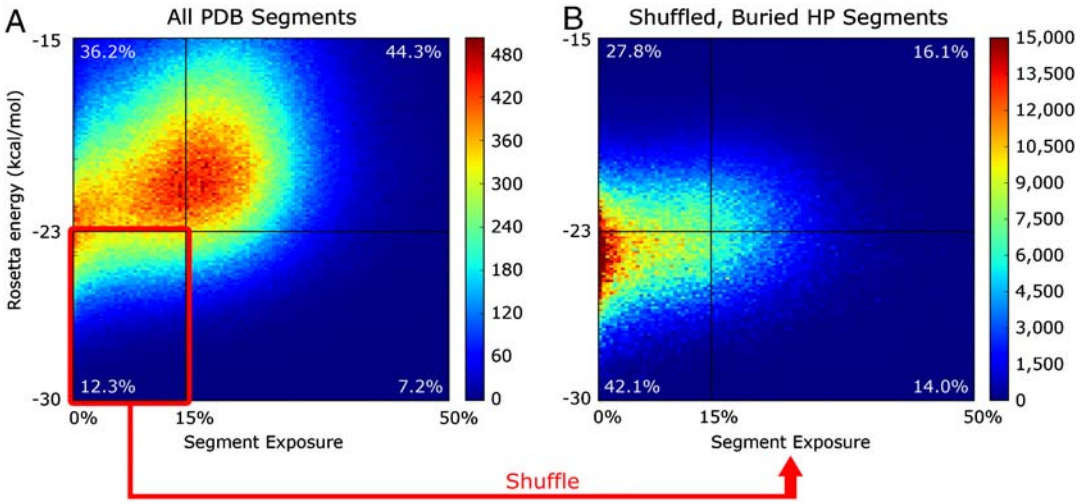

Fig. 2. Sequence is more important than residue composition in determining propensity for formation of amyloid-like fibrils. (*A*) shows on the *y* axis the fibrillation propensity, calculated by the 3D profile method, and on the *x* axis the solvent exposure of segments in the nonredundant structures (proteins having less than 50% sequence identity) in the PDB. Segments below the horizontal −23 kcal/mol line are HP and are preferentially buried if <15% exposed (vertical line). Thus buried HP segments are plotted in the lower left quadrant and are enclosed by a red rectangle. In (*B*) these segments are shuffled, and the fibrillation propensity and exposure of the resulting shuffled segments are plotted. The shuffling operation, which changes the residue order within a segment but does not alter its residue composition, allows these segments to migrate from the lower left quadrant to lower values of propensity (higher energy) and greater exposure to solvent. Only 42% of segments remain buried and retain their high fibrillogenic propensity. The exposure to solvent of these segments is estimated by examining the corresponding positions within proteins that contain the exact residue sequence. Of the 720 permutations for a given six-residue segment, about 12 are found in some PDB protein, and hence the solvent exposure can be determined and plotted. Warmer colors indicate higher count density. The percentage of segments found in each quadrant is indicated in white. The implications are that high propensity depends on a nonrandom pattern of sequences and that proteins tend to bury HP segments.

(<0.10%) surface beta strands with high propensity available to form an intermolecular steric zipper. We conclude that protein folds have evolved to remove segments of high propensity and proper conformation for fibrillation from protein surfaces. Further reinforcing this finding is our observation that HP segments are one-third less likely to be found in natively disordered segments of protein chains, where they might more easily interact with HP segments from identical molecules.

We find that HP segments tend to be buried or twisted into unfavorable conformations for forming beta sheets, which is consistent with the finding of Pertinhez (7) and others (1, 8–10, 24, 25) that destabilization of proteins can lead to fibrillation. For some proteins a delicate balance between protein folding and misfolding exists that can be tipped by changes in environment, destabilizing mutations, or even protein concentration. As a protein undergoes partial denaturation, HP segments are exposed, such that nucleation or elongation of fibrils can take place with the newly exposed HP segments (17).

The importance of HP segments in forcing fibrillation is in accord with previous studies (18, 26), where we have forced RNase A into fibril form by inserting HP segments into the C-terminal hinge loop, after residue 112. This position is marked by the * in the structure in Fig. 1. These inserts include HP segments from proteins capable of forming amyloid fibrils such as A-beta. Insertion of each exogenous segment into the flexible C-terminal hinge loop allows the segment to form intermolecular steric zipper interactions with an identical segment. This interaction is sufficient to drive nonfibrillizing RNase A to form fibrils. Whether or not an HP segment enables a protein to belong to the amylome depends on the position of the segment within the sequence. Proteins may have evolved to permit such segments only in positions that do not lead to fibrils that interfere with normal function. In the case of RNase A, neither the endogenous HP segments nor the native C-terminal hinge loop drive RNase A to form fibrils. Furthermore, conformational flexibility of a segment, governed by adjacent segments of the protein chain, is a key determinant of fibrillation propensity. The insertion of HP segments into the C-terminal hinge loop of RNase A apparently permits them to participate in steric zippers, whereas the more limited flexibility of endogenous HP segments within the wild-type RNase A sequence does not. Conversely, the insertion of a segment with high energy and thus low propensity for fibrillation at the same position in the C-terminal hinge loop does not convert the protein into a fibril former. RNase A is an example of a protein whose sequence seems to self-chaperone itself, not permitting its HP segments to form intermolecular steric zippers. In other globular proteins, such as lysozyme (27), superoxide dismutase (28), or transthyretin (25), misfolding or destabilization of the native structure can lead to exposure of HP segments, which then progresses to fibrillation.

In addition to the self-chaperoning effects described above, proteins are also protected from fibrillation during the process of folding by molecular chaperones (29). In a first line of defense against fibrillation, a chaperone such as Hsp70 sequesters protein chains, allowing them to fold without interference of identical chains, which not only reduces the possibility of 3D domain swapping, but also slows fibrillation that might otherwise occur through steric zipper formation at HP segments.

What factors enable a protein to belong to the amylome? Our results demonstrate that eligible proteins must contain at least one segment with HP and sufficient conformational freedom to form an intermolecular steric zipper. The amino acid composition of the segment is not necessarily indicative of its propensity to form a steric zipper; rather, a specific residue sequence within the segment is required. We find that those self-complementary segments are common but, in folded proteins, are constrained from interaction with each other. When such segments become exposed by partial denaturation or in artificial constructs, the protein can be converted to a fibril former. Therefore, both the presence of an HP segment in the protein and sufficient accessibility to such a segment are necessary to enable a protein to belong to the amylome. Experiments investigating the effect of accessibility of HP segments will be the next milestone in understanding why proteins form fibrils.

Methods

3D Profile Method. Our fibrillogenic predictions by the 3D profile method are available online in the ZipperDB database at http://services.mbi.ucla.edu/zipperdb. The database is easily searchable by protein identifier, name, or sequence. For each protein, we provide the fibrillogenic propensity profile and, for the hexapeptide segments, Rosetta energies and our steric zipper models. Users will be able to submit their own protein sequences for analysis in the near future.

Datasets. *E. coli*, *S. cerevisiae*, and *H. sapiens* genomes were retrieved from the National Center for Biotechnology Information Web site. A nonredundant set of the PDB based on 50% sequence identity containing 12,846 proteins was retrieved from the PDB in December 2007.

Fibrillation Propensity Predictions. Fibrillation propensities were computed for all proline-free six-residue segments in the protein sequences by using the 3D profile method (14). To avoid problems with their disulphide bonding abilities, cysteines were substituted to serines during modeling. Briefly, we used the backbone coordinates from the crystal structure of NNQQNY as a template and evaluated the energetic fit of a putative sequence by "threading" it onto the template with RosettaDesign (20). Segments that have a low Rosetta energy (−23 kcal/mol or lower) and a steric zipper-like interface, as judged by its shape complementarity (>0.7), are classified to have HP.

For larger datasets, we estimated the propensity by using a simplified "triplet" method, which takes into account only the interactions of the three inward-facing residues at the steric zipper interface (residues 1, 3, and 5 of each hexapeptide). The three outward-facing residues are modeled as alanines, which significantly reduces computational complexity at the expense of accuracy. We find that this method introduces an average error of 4 kcal/mol per hexapeptide and has the tendency to overestimate the propensity by calculating a lower energy than the 3D profile method for a given hexapeptide. See *SI Appendix*.

Aggregation Propensity Predictions with TANGO. Aggregation propensity of peptide segments was predicted with TANGO (11) by using default parameters. For segments that originated from natural protein sequences, the full length wild-type protein sequence was used; for designed control segments, a construct consisting of the segment of interest flanked by six glycines on both sides was used. We classified the segment as high propensity if TANGO assigned a score of ≥ 1 for any residue within the segment, which is a very liberal threshold because TANGO reported peak propensities of 93 (A-beta), 97 (beta microglobulin), 4.5 (superoxide dismutase), and 2.5 (islet amyloid polypeptide) in our benchmarks.

Segment Exposure in Protein Structures. Segment exposure was calculated from the protein structure containing the said segment by using the solvent accessible surface area reported by DSSP (30). The residue exposure reported here is equal to the solvent accessibility of the residue normalized by the maximum solvent accessibility of the corresponding residue type free in space. The segment exposure is the average exposure of the residues contained therein. Exposed segments are those whose exposure is at least equal to the median exposure over the entire set, which was determined to be 15% (*SI Appendix*).

Segment Geometry. The alpha carbon rmsd of each hexapeptide to the NNQQNY backbone was computed with the Kabsch algorithm (31). Segments with rmsd <0.75 Å are deemed to have compatible beta-strand geometry.

Functional and Conserved Residues. The UniProt database was queried for active site residue annotations in proteins. Segments containing one or more such active site residues were classified as functional. Separately, a multiple sequence alignment was prepared with BLAST from the National Center for Biotechnology Information nonredundant gene database, by using an expectation value threshold of 1×10^{-10}. Pairwise similarity scores were computed from the alignment by using the BLOSUM62 matrix to rank conserved residues within a protein candidate; various conservation rank thresholds (such as top 5 or 10 residues) were used to identify conserved residues.

Segment Shuffling. Given a hexapeptide segment, all 720 (6!) sequence permutations were enumerated. This operation rearranges the residues within a segment while keeping its composition constant. For any rearranged sequence that was present in any protein in the dataset, the fibrillation propensity was computed by using the 3D profile method as described above. The exposure of the permuted segment was computed from the corresponding structure.

Fibril Formation and Electron Microscopy. Lyophilized, synthetic peptides (CSBio) were dissolved in various buffer, solvent, and salt solutions and incubated at 37 °C with vigorous shaking. Samples were applied directly to hydrophillic 400-mesh carbon-coated formvar support films mounted on copper grids (Ted Pella). After 2.5 min, the grids were rinsed with 0.2 μm-filtered distilled water and stained for 1 min with 1% uranyl acetate. Dried grids were examined with a Philips CM120 transmission electron microscope at an acceleration voltage of 120 keV.

1. Astbury WT, Dickinson S, Bailey K (1935) The x-ray interpretation of the denaturation and the structure of the seed globulihns. *Biochem J* 29:2351–2360.
2. Terry WD, et al. (1973) Structural identity of Bence Jones and amyloid fibril proteins in a patient with plasma cell dyscrasia and amyloidosis. *J Clin Invest* 52:1276–1281.
3. Sunde M, et al. (1997) Common core structure of amyloid fibrils by synchrotron x-ray diffraction. *J Mol Biol* 273:729–739.
4. Geddes AJ, Parker KD, Atkins ED, Beighton E (1968) "Cross-beta" conformation in proteins. *J Mol Biol* 32:343–358.
5. Kelly J, Balch WE (2003) Amyloid as a natural product. *J Cell Biol* 161:461–462.
6. Maji SK, et al. (2009) Functional amyloids as natural storage of peptide hormones in pituitary secretory granules. *Science* 325:328–332.
7. Pertinhez TA, et al. (2001) Amyloid fibril formation by a helical cytochrome. *FEBS Lett* 495:184–186.
8. Fink AL (1998) Protein aggregation: Folding aggregates, inclusion bodies and amyloid. *Fold Des* 3:R9–23.
9. Sunde M, Blake CC (1998) From the globular to the fibrous state: Protein structure and structural conversion in amyloid formation. *Q Rev Biophys* 31:1–39.
10. Chiti F, Dobson CM (2009) Amyloid formation by globular proteins under native conditions. *Nat Chem Biol* 5:15–22.
11. Fernandez-Escamilla A-M, Rousseau F, Schymkowitz J, Serrano L (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol* 22:1302–1306.
12. Tartaglia GG, et al. (2008) Prediction of aggregation-prone regions in structured proteins. *J Mol Biol* 380:425–436.
13. Trovato A, Chiti F, Maritan A, Seno F (2006) Insight into the structure of amyloid fibrils from the analysis of globular proteins. *PLoS Comput Biol* 2:e170.
14. Thompson MJ, et al. (2006) The 3D profile methods for identifying fibril-forming segments of proteins. *Proc Natl Acad Sci USA* 103:4074–4078.
15. Dobson CM (2004) *Seminars in Cell and Developmental Biology*, ed Ellis J Vol 15 (Associated Press, New York), pp 3–16.
16. Balbirnie M, Grothe R, Eisenberg D (2001) Atomic structures of amyloid cross-beta spines reveal varied steric zippers. *Proc Natl Acad Sci USA* 98:2375–2380.
17. Nelson R, et al. (2005) Structure of the cross-beta spine of amyloid-like fibrils. *Nature* 435:773–778.
18. Sambashivan S, Liu Y, Sawaya MR, Gingery M, Eisenberg D (2005) Amyloid-like fibrils of ribonuclease A with three-dimensional domain-swapped and native-like structure. *Nature* 437:266–269.
19. Sawaya MR, et al. (2007) Atomic structures of amyloid cross-beta spines reveal varied steric zippers. *Nature* 447:453–457.
20. Kuhlman B, Baker D (2000) Native protein sequences are close to optimal for their structures. *Proc Natl Acad Sci USA* 97:10383–10388.
21. Rousseau F, Serrano L, Schymkowitz JW (2006) How evolutionary pressure against protein aggregation shaped chaperone specificity. *J Mol Biol* 355(5):1037–1047.
22. Ivanova MI, Thompson MJ, Eisenberg D (2006) A systematic screen of beta(2)-microglobulin and insulin for amyloid-like segments. *Proc Natl Acad Sci USA* 103:4079–4082.
23. Wiltzius JJW, et al. (2009) Molecular mechanisms for protein-encoded inheritance. *Nat Struct Mol Biol* 16:973–978.
24. Hurle MR, Helms LR, Li L, Chan W, Wetzel R (1994) A role for destabilizing amino acid replacements in light-chain amyloidosis. *Proc Natl Acad Sci USA* 91:5446–5450.
25. Lai Z, Colon W, Kelly JW (1996) The acid-mediated denaturation pathway of transthyretin yields a conformational intermediate that can self-assemble into amyloid. *Biochemistry* 35:6470–6482.
26. Teng PK, Eisenberg D (2009) Short protein segments can drive a non-fibrillizing protein into the amyloid state. *Protein Eng Des Sel* 22:531–536.
27. Pepys MB, et al. (1993) Human lysozyme gene mutations cause hereditary systemic amyloidosis. *Nature* 362:553–557.
28. Elam JS, et al. (2003) An alternative mechanism of bicarbonate-mediated peroxidation by copper-zinc superoxide dismutase: rates enhanced via proposed enzymeassociated peroxycarbonate intermediate. *J Biol Chem* 278:21032–21039.
29. Raman B, et al. (2005) AlphaB-crystallin, a small heat-shock protein, prevents the amyloid fibril growth of an amyloid beta-peptide and beta2-microglobulin. *Biochem J* 392:573–581.
30. Kabsch W, Sander C (1983) Identical pentapeptides with different backbones. *Biopolymers* 22:2577–2637.
31. Kabsch W (1976) A solution of the best rotation to relate two sets of vectors. *Acta Crystallogr* A32:922–923.