Detecting hidden sequence propensity for amyloid fibril formation

SUKJOON YOON¹ AND WILLIAM J. WELSH

Department of Pharmacology, University of Medicine and Dentistry of New Jersey–Robert Wood Johnson Medical School, Piscataway, New Jersey 08854, USA

(RECEIVED April 5, 2004; FINAL REVISION May 10, 2004; ACCEPTED May 18, 2004)

Abstract

The preponderance of evidence implicates protein misfolding in many unrelated human diseases. In all cases, normal correctly folded proteins transform from their proper native structure into an abnormal β -rich structure known as amyloid fibril. Here we introduce a computational algorithm to detect nonnative (hidden) sequence propensity for amyloid fibril formation. Analyzing sequence–structure relationships in terms of tertiary contact (TC), we find that the hidden β -strand propensity of a query local sequence can be quantitatively estimated from the secondary structure preferences of template sequences of known secondary structure found in regions of high TC. The present method correctly pinpoints the minimal peptide fragment shown experimentally as the likely local mediator of amyloid fibril formation in β -amyloid peptide, islet amyloid polypeptide (hIAPP), α -synuclein, and human acetylcholinesterase (AChE). It also found previously unrecognized β -strand propensities in the prototypical helical protein myoglobin that has been reported as amyloidogenic. Analysis of 2358 nonhomologous protein domains provides compelling evidence that most proteins contain sequences with significant hidden β -strand propensity. The present method may find utility in many medically relevant applications, such as the engineering of protein sequences and the discovery of therapeutic agents that specifically target these sequences for the prevention and treatment of amyloid diseases.

Keywords: amyloid fibril; tertiary contacts; secondary structure; hidden β -strand propensity; H β P, SCOP

Amyloid fibril formation, or amyloidosis, a manifestation of protein misfolding, is widely observed in many unrelated human diseases, including common neurodegenerative and neuromuscular pathologies such as Alzheimer's disease, Parkinson's disease, and Huntington's disease (Sacchettini

and Kelly 2002). It also figures prominently in type II (or late-onset) diabetes mellitus. The clinical manifestations of amyloidosis are varied and depend on the biochemical nature of the fibril protein, as well as the area of the body or organ that is involved. Amyloid fibril formation is also associated with the so-called prion diseases, known as spongiform encephalopathies, and most mammalian species are believed susceptible to prion diseases. Recognition of those factors that influence protein conformational changes and misfolding is thus an urgent priority in elucidating the fundamental causes of, and finding effective treatments for, these afflictions. Unfortunately, a definitive explanation of the root cause and molecular mechanism of amyloid fibril formation remains elusive. Likewise, no reliable computational method exists for predicting propensities toward amyloid fibril formation. We believe that the present work makes a serious contribution toward addressing this latter need.

Reprint requests to: William J. Welsh, Department of Pharmacology, University of Medicine & Dentistry of New Jersey–Robert Wood Johnson Medical School, Piscataway, NJ 08854, USA; e-mail: welshwj@umdnj.edu; fax: (732) 235-3475.

¹Present address: ArQule, Inc., Woburn, MA 01801, USA.

Abbreviations: A β , β -amyloid; AChE, human acetylcholinesterase; ANN, artificial neural network; BuChE, butyrylcholinesterase; DSSP, definition of secondary structure of proteins; H β P, hidden β propensity; hIAPP, human islet amyloid precursor protein; NAC, non-A β component of Alzheimer's disease amyloid; PAM, percentage accepted mutations; PDB, Protein Data Bank; SCOP, structural classification of proteins; TC, tertiary contact.

Article and publication date are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.04790604>.

As the conformation of a local amino-acid sequence can be influenced by its tertiary environment (Minor and Kim 1996), consideration of the tertiary context of a sequence can improve our understanding of sequence–structure relationships in local regions of proteins. It is possible to quantify the influence of tertiary context by a simple approach that counts the number of atom-to-atom tertiary contacts (TCs), rather than resorting to exhaustive energy calculations (Berezovsky and Trifonov 2001). TCs are formed between nonadjacent residues when a protein undergoes three-dimensional folding, bringing together residues that can be far apart along the linear amino-acid sequence. The notion of TCs has been applied successively by other workers for the identification of self-stabilizing folding units (Fischer and Marqusee 2000) and for comparisons of protein-folding complexity and kinetics (Plaxco et al. 1998). Counting TCs between nonbonded atoms provides a simple yet effective way to approximate tertiary interactions and solvent accessibility, thus representing a judicious compromise between speed and rigor suitable for rapid predictions on a large scale. Here we introduce a computational algorithm that detects the nonnative (hidden) β -strand propensity (H β P) of sequences by formulating relationships between protein local sequence and secondary structure in terms of TCs. This algorithm is henceforth called the H β P method.

Algorithms for making predictions of protein native secondary structures typically rely upon associating homologies between the query sequence and template sequences for which the three-dimensional structure is known. These algorithms generally require a minimum sequence context of 14 to 17 residues to determine the unique native secondary structure of a query peptide or protein (Pan et al. 1999), whereas the present H β P method is applied by using a much shorter sequence context (seven residues), which is more likely to retrieve multiple template secondary structures for a given query sequence (Zhou et al. 2000). Essentially, the present method partitions the structural determinants for local secondary structure propensity into two independent variables: (1) local effects (flanking sequences) and (2) non-local effects (TCs). The sequence similarity of flanking regions is an efficient measure of local effects on secondary structure propensity. A seven-residue sequence context was chosen in part because it represents the minimum size sufficient to account for possible ($i, i + 3$) side-chain interactions within helical sequences. In light of the uncertainty of the homology between similar or even identical seven-residue sequences within the context of evolutionary relationships (Zhou et al. 2000), an obvious advantage in using a shorter (e.g., seven-residue) rather than a longer sequence context when searching protein fold and/or structural databases is the likelihood that a greater number of similar sequences in nonhomologous proteins (e.g., diverse TC states) will be retrieved. By structural analysis of this larger pool of similar sequences, it becomes possible to systematically

evaluate sequence–structure relationships in terms of TCs (Fig. 1). We have implemented this scheme in a computational procedure that predicts the nonnative secondary structure preferences of a query local sequence by searching the SCOP20 (Structural Classification of Proteins) database (Brenner et al. 2000) for conformational preferences of similar local sequences that vary with respect to their TC states. By capturing the influence of variations in tertiary structural environment on local conformations, the present H β P method pinpoints those regions in a protein that exhibit high (or low) propensity for undergoing conformational change. In this report, we illustrate how this method is applied to detect the H β P in protein fragments known to be associated with amyloid fibril formation. The present H β P algorithm is not intended to ascertain whether a specific protein is amyloidogenic; however, it will detect sequences within the protein that are conducive to triggering amyloid fibril formation (i.e., strong H β P).

Figure 1. Flowchart of our TC-based scheme for calculating secondary structure propensities $P(\alpha|low)$ and $P(\beta|high)$. The propensities of α -helix in low TC bins, $P(\alpha|low)_{temp}$, and β -strand in high TC bins, $P(\beta|high)_{temp}$, were calculated from the number of templates in low and high TC bins, respectively. The query sequence GEAVELA is used as an example (for details, see Materials and Methods). The propensities of a query are predicted from the corresponding correlation functions f^* and g^* of its templates taken from the correlation curves in Figure 6.

Amyloid fibril formation, an intriguing example of protein conformational changes, is associated with an increase in β -strands, leading to fibrillar aggregation (Jimenez et al. 1999). Recent studies have shown that diverse proteins not related to amyloid disease can also aggregate into fibrils under laboratory-controlled destabilizing conditions (Chiti et al. 1999; Fändrich et al. 2001). Although amyloid fibrils share a common core of highly compact cross- β structure (Baltirnie et al. 2001; Fändrich et al. 2001), the lack of consensus sequence among amyloidogenic proteins, together with the recent observation of helical content within amyloid fibrils (Mangione et al. 2001), suggests a more varied secondary structure. Encouragingly, investigators have recently been able to establish some correlation between sequence conservation and amyloid fibril formation by conducting a large-scale sequence-structure analysis (Benyamini et al. 2003a,b). Identification of those sequences in a protein that exhibit a strong propensity for nonnative β -strand formation would help scientists to decipher the apparently complex interplay of secondary structure and conformational features that may trigger amyloid

fibril formation in virtually any protein. Secondary structure prediction methods, such as the popular PHD algorithm (Rost 1996), were specifically designed for the intended purpose of predicting native secondary structure based on sequence homology. Thus, although demonstrating some proficiency in detecting nonnative β -structure in at least one study (Kallberg et al. 2001), PHD predicts virtually zero β -strand propensity for myoglobin (Fig. 2A). Nevertheless, this prototypical α -helical globular protein has been shown to form amyloid fibrils under denaturing conditions (Fändrich et al. 2001). Thus far, no secondary-structure prediction method exists that will detect H β P for amyloid fibril formation in common globular protein sequences. H β P indicates the tendency of some sequences that, although α -helix or random coil in the native state, can transform to nonnative β -strands under certain conditions that are nonetheless physiologically relevant. Hence, our primary objective was to develop a sensitive measure of H β P by calculating TC-based secondary structure. This knowledge bears relevance to medically relevant applications, such as the engineering of proteins devoid of such sequences and the

(A) PHD prediction

```

| DGEWQQVLNVWGKVEADIAGHGQEV LIRLFTGHPETLEKFDKFKHLKTEA |
PHD | HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH |
prH | 799999999999999999999999999999999999999999999999999 |
prE | 000000000000000000000000000000000000000000000000000 |

| EMKASEDLKKHGTVVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKY |
PHD | HH HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH |
prH | 554216678878999999999999999999999999999999999999999 |
prE | 000000000000000000000000000000000000000000000000000 |

| LEFISDAI I HVLHSHKHPGDFGADAQGAMTKALELFRNDIAAKYKELG |
PHD | HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH |
prH | 899999999999999999999999999999999999999999999999999 |
prE | 000000000000000000000000000000000000000000000000010 |
    
```

α β Probability

0.95 ~	0.95 ~
0.80 - 0.95	0.80 - 0.95
0.65 - 0.80	0.65 - 0.80
0.50 - 0.65	0.50 - 0.65

(B) H β P method

Figure 2. Secondary structure propensity of the horse muscle myoglobin sequence. Red-colored letters represent α -helical regions in the native structure. (A) Secondary structure propensity as predicted by PHD method. PHD-generated propensities for helix (prH) and extended β -strand (prE) conformations are presented numerically using a 0-to-9 scale, in which “0” indicates zero to very low probability and “9” indicates very high probability to near certainty. The PHD-generated propensity for random coil is not shown. The letter “H” indicates that PHD predicted a helical native secondary structure from among three possible states (helix, β -strand, and coil). (B) The TC-based propensities, $P(\alpha/low)$ and $P(\beta/high)$, using the present H β P method. The boxes indicate sequences predicted to possess strong nonnative β -strand propensity.

discovery of therapeutic agents that specifically target these regions associated with fibrillar aggregation.

Results and Discussion

TC-dependent secondary structure propensities

To understand the correlation between secondary structure and TCs in local regions of a protein, the relative occurrence of secondary structure elements in the SCOP20 database was analyzed at various levels of TCs (Table 1). Although random coil occurs most frequently across the entire range of TCs, the occurrence of α -helix and β -strand shows opposing trends at low and high TCs. These associations between TCs and secondary structure elements (α -helix and β -strand) suggest that TC filters can improve our ability to predict relationships between sequence and structure in local regions. Previous studies (Pan et al. 1999; Zhou et al. 2000) have suggested that a seven-residue sequence context is too short for the prediction of native secondary structure. Nevertheless, we found strong correlations between query and template sequences within a seven-residue context when TC filters are used to evaluate the TC-dependent secondary structure propensity of a given query sequence (Table 2).

The predictive accuracy shown in Table 2 was found by analyzing SCOP20 fragments found with low or high TCs as test queries. The prediction accuracy for α -helix in low TCs was 75%, using $P(\alpha|low)_{query} > 0.5$ as the selection criterion. This criterion provides 92% coverage of all helical fragments with low TCs in SCOP20. The prediction accuracy and percentage of coverage are inversely related and dependent upon the criterion chosen for $P(\alpha|low)_{query}$ and $P(\beta|high)_{query}$. For example, the more stringent criterion $P(\alpha|low)_{query} > 0.9$ yields 90% prediction accuracy but lower coverage (19%). Similarly, the prediction accuracy for β -strand in high TCs was 75% using $P(\beta|high)_{query} > 0.5$ as the selection criterion. This criterion provides 66% coverage of all β -strand fragments with high TCs in SCOP20. The prediction accuracy and coverage are generally lower for β -strand in high TCs than for α -helix in low

TCs. This result is also commonly found for native secondary structure prediction methods. Because β -strands occur less frequently than do α -helices in proteins (Table 1), it is more difficult to predict the β -strand propensity than α -helical propensity. In effect, the size of template pool is smaller for β -strand than for α -helix. This limitation is not a factor when identifying local regions that exhibit high H β P, because it is unnecessary to assign secondary structures to all residues in a sequence. In summary, validation of the present H β P method was conducted as described above by using the SCOP20 fragments to predict their native secondary structures in their native TC states. This method was then applied to predict the nonnative secondary structure propensities in nonnative TC states (whether “high” or “low”).

Because the data set of 453,787 amino-acid fragments were extracted from nonhomologous SCOP20 domains, the 30 top-scoring templates for a given query sequence will likely represent diverse TC states. Inasmuch as the seven-residue templates are rarely found exclusively in low or high TC bins, a given query sequence will typically yield nonzero values for both $P(\alpha|low)$ and $P(\beta|high)$. In contrast to other computational methods that predict the native secondary structure of a protein sequence, the present H β P method was designed to predict propensities for nonnative secondary structure. Its intended purpose is to pinpoint local regions that are more susceptible to conformational change, for example, from helical to β -strand. Given the preponderance of evidence that increased β -strand formation is a common feature in triggering aggregation of proteins into amyloid fibrils (Chiti et al. 2003; Paz and Serrano 2004), the propensity for amyloid fibril formation can be predicted by examining the H β P in high TC environments (i.e., $P[\beta|high]$) for sequences in which β -strand is not the native structure.

Pinpointing amyloid fibril-forming sequences

The wild-type β -amyloid peptide (A β) is well known for its strong propensity to form fibrils and, as such, serves as a highly relevant test case for the present method. Tjernberg et al. (1996) deduced that the KLVFF segment in the truncated native A β sequence (N-terminal residues 1–18) was of critical importance in the polymerization of amyloid fibril, whereas a mutant sequence in which KLVFF was replaced by AAVFA showed a markedly reduced tendency to form amyloid fibrils. Their work provides convincing evidence that even short sequences may govern a the propensity of a protein for amyloid fibril formation. We calculated $P(\alpha|low)$ and $P(\beta|high)$ for the wild-type A β (N-terminal residues 1–28) and for the corresponding AAVFA mutant sequence. Consistent with the findings of Tjernberg et al. (1996), the present H β P method predicted that the KLVFF segment in A β shows the strongest nonnative β -strand propensity in high TCs, whereas the alternative alanine-rich

Table 1. The relative occurrences of secondary structure elements in different TC states from surveying 453,787 fragments, each seven residues in length, from sequences in the SCOP20 database

Secondary structure	Coil	α	β	Total sequences
Tertiary contacts				
Low	38%	59%	3%	191,300
Intermediate	47%	37%	16%	112,199
High	39%	11%	50%	150,288
All	41%	38%	21%	453,787

Table 2. Prediction accuracy of the present H β P method

	<i>P</i> (α low)			<i>P</i> (β high)		
	No. correct predictions	% prediction accuracy	% helix coverage	No. correct predictions	% prediction accuracy	% β -strand coverage
>0.9	21,985	90	19	8,294	85	11
>0.7	71,062	82	63	31,710	78	42
>0.5	94,989	75	92	49,404	70	66

Every seven-residue fragment from SCOP20 with low or high TCs was used as a query once. The propensity of α -helix in low-TC bins, *P*(α |low), or β -strand in high-TC bins, *P*(β |high), for the query is calculated from the fitted curved in Fig. 6, A or B. PAM30 cutoff score = 15. % helix coverage is the ratio of correctly predicted α -helices vs. total α -helices; % β -strand coverage is the ratio of correctly predicted β -strands vs. total β -strands.

AAVFA segment shows substantially reduced β -strand propensity (Fig. 3A,B, respectively). The PHD method, although also predicting β -strand for the KLVFF sequence in A β , incorrectly, predicts the native secondary structure of A β (Fig. 3A). Our H β P method, which calculates *P*(α |low) and *P*(β |high) independently, predicts strong helical propensity at low TCs in accordance with the native structure of A β in the solution state.

The islet amyloid polypeptide (hIAPP) has been shown to accumulate as amyloid fibrils in the pancreas of individuals with type II diabetes (Hoppener et al. 2000). By using systematic experimental approaches, Mazor et al. (2002) identified the major recognition motif within hIAPP capable of self-assembly. They first examined the ability of hIAPP to bind with 28 overlapping peptides (decamers) that span the entire hIAPP sequence. Comparison of their experimental binding data for the 28 decamers (taken from Fig. 3b in Mazor et al. 2002) against predicted β propensity revealed a strong correlation with that calculated by the present H β P method ($r = 0.6$) but not with that calculated by PHD ($r = -0.2$; Table 3, Fig. 4). The highest H β P values calculated by the present method were associated with fragments 11 and 12 (RLANFLVHSS and LANFLVHSSN, respectively), in agreement with the results obtained by Mazor et al. (2002). In contrast, the PHD method failed to detect any β propensity (prE) in this major motif. Both the present H β P method and PHD predicted some β propensity in two minor motifs (fragments 2, 19–21). Two overlapping pentapeptides (NFLVH and FLVHS) within the major binding domain of hIAPP_{11–20} (RLANFLVHSS, fragment 11 in Table 3) were determined by Mazor et al. (2002) as the shortest active fragments capable of self-association. Consistent with these experimental findings, our H β P method pinpointed the NFLVH region as possessing the highest H β P (i.e., highest *P*(β |high) value) in the entire hIAPP sequence (Fig. 3C). The consistency of the present H β P method with these experimental results for hIAPP lends credibility to the sensitivity of our TC-based approach in correctly predicting non-native β -strand propensity even in extremely short sequences. Identification of short recognition motifs, such as

NFLVH in hIAPP, is of critical value in efforts to discover “ β -sheet breakers,” much like AcQKL β FFNH₂ was shown by Tjernberg et al. (1996) to halt fibrillization of A β .

The major fibrillar material of Parkinson’s disease was shown to be α -synuclein (Spillantini et al. 1997, 1998). Interestingly, the peptide derived from the central hydrophobic region of α -synuclein represents a second major intrinsic constituent of Alzheimer’s plaques. This 35-amino-acid peptide, known as NAC (non-A β component of Alzheimer’s disease amyloid) was shown to constitute about 10% of the amyloid plaque (Ueda et al. 1993). The amyloidogenicity is not uniformly distributed within NAC. For example, the C-terminal half of the peptide (NAC residues 19–35, QKTVEGAGSIAAATGFV) does not fibrillate, whereas the N-terminal fragment 3–18 (VTNVGGAVVT GVTAVA) can fibrillate (El-Agnaf et al. 1998; Bodles et al. 2001). NAC fragment 11–22 (VTGVTAVAQKTV) and 8–18 (GAVVTGVTAVA) also showed propensity to fibrillate (Bodles et al. 2001; Giasson et al. 2001). Consistent with this experimental evidence, our H β P method detected significant H β P exclusively in the N-terminal region (Fig. 3D). The overlapping core (VTGVTAVA) of the above three fibril-forming fragments was predicted to possess exceptionally strong β -strand propensity. In sharp contrast with experimental evidence and the H β P method, PHD predicts higher β -strand propensity in the C-terminal region and a mixture of α and β propensities in the core region (VTGVTAVA).

The intact human acetylcholinesterase (AChE) C-terminal domain is α -helical in the native state, but a shorter, 14-residue fragment (AChE_{586–599}) forms β -rich amyloid fibrils (Cottingham et al. 2003). These fibrous AChE_{586–599} aggregates possess all the classical hallmarks of amyloid fibrils and are neurotoxic in vitro (Cottingham et al. 2002). However, the fragment (BuChE_{573–586}) derived from the closely related enzyme butyrylcholinesterase (BuChE) showed no detectable fibril formation (Cottingham et al. 2003). Consistent with these observations, the H β P method predicts significantly greater β -strand propensity for AChE_{586–599} than for BuChE_{573–586} (Fig. 3E,F). The PHD

Figure 3. $P(\alpha/low)$ and $P(\beta/high)$ profiles of amyloidogenic sequences calculated by the present HbP method, compared with secondary structure propensity predicted by the PHD method. PHD-generated propensities for prH and prE are presented similarly as in Figure 2. (A) The wild-type Aβ₄₋₂₅ peptide associated with amyloid fibril formation in Alzheimer's disease. The residues enclosed in the box indicate five key residues (KLVFF) essential for fibrillar polymerization (Tjernberg et al. 1996). Using standard single-letter codes, residue names shaded in red adopt α -helical conformations in the soluble form. (B) Corresponding mutant Aβ fragment with three alanine substitutions that exhibits reduced tendency to form amyloid fibrils (Tjernberg et al. 1996). Boldface letters represent amino-acid substitutions. (C) The hIAPP₄₋₃₄ sequence associated with amyloid fibril formation in type II diabetes. The residues enclosed in the box indicate two overlapped five residues (NFLVH and FLVHS) that are capable of self-association and thus may serve as the core molecular recognition sequence for amyloid fibril formation (Mazor et al. 2002). (D) NAC sequence of α -synuclein. The residues enclosed in the box indicate the overlapping sequence in three truncated fragments showing high amyloidogenic propensity (for details, see text). (E) Amyloidogenic AChE₅₈₆₋₅₉₉ fragment. (F) Nonamyloidogenic BuChE₅₇₃₋₅₉₆ fragment.

method predicts a helical secondary structure in both fragments and was unable to detect any β -strand propensity in AChE₅₈₆₋₅₉₉. The ability of the HbP method to differentiate between these two highly similar sequences (i.e., AChE₅₈₆₋₅₉₉ and BuChE₅₇₃₋₅₈₆) with respect to their reported propensity to form amyloid fibrils (Cottingham et al. 2003) speaks to its exceptional sensitivity.

A summary of results collected for known amyloidogenic and nonamyloidogenic subsequences is assembled in Table 4 (below). $P(\beta/high)$ is ≥ 0.5 for all amyloidogenic cases, whereas only $P(\beta/high) = 0.2$ for the two nonamyloidogenic sequences (i.e., NAC sequence 19–35 of α -synuclein; BuChE₅₇₃₋₅₉₆). The $P(\beta/high)$ level of each sequence calculated by the present HbP method is consistent with cor-

Table 3. Comparison of predicted β propensity and experimentally determined binding ability of local sequences to hIAPP

	Sequences fragment	Binding	prE ^a	H β P ^b
1	KCNTATCATQ	0	0.1	0.5
2	CNTATCATQR	40	0.1	0.6
3	NTATCATQRL	0	0	0
4	TATCATQRLA	0	0	0
5	ATCATQRLAN	0	0	0
6	TCATQRLANF	0	0	0
7	CATQRLANFL	32	0	0
8	ATQRLANFLV	39	0	0
9	TQRLANFLVH	30	0	0.6
10	QRLANFLVHS	60	0	0.8
11	RLANFLVHSS	100	0	0.9
12	LANFLVHSSN	82	0	0.9
13	ANFLVHSSNN	0	0	0.8
14	NFLVHSSNNF	0	0	0
15	FLVHSSNNFG	0	0	0
16	LVHSSNNFGA	0	0	0
17	VHSSNNFGAI	0	0	0
18	HSSNNFGAIL	0	0.1	0
19	SSNNFGAILS	8	0.1	0
20	SNNFGAILSS	10	0.1	0.6
21	NNFGAILSST	5	0.5	0.7
22	NFGAILSSTN	0	0.6	0.7
23	FGAILSSTNV	0	0.5	0.6
24	GAILSSTNVG	0	0.3	0
25	AILSSTNVGS	0	0	0
26	ILSSTNVGSN	0	0	0
27	LSSTNVGSNT	0	0	0
28	SSTNVGSNTY	0	0.1	0
Correlation to binding exp. (<i>r</i>)			-0.2	0.6

^a β -Strand propensity predicted by PHD method.^b Hidden β -strand propensity (H β P) predicted by present method.

All propensities are residue averages. Boldface letter refers to the residue for which values of prE (from PHD) and H β P (from present method) are listed on the same line. Binding data for the 28 decamers were taken directly from Fig. 3b in Mazor et al. (2002). The graphical details of propensities are shown in Fig. 3C.

responding experimental observations. In contrast, the PHD method generally predicted extremely low β propensity in amyloidogenic fragments except in the A β sequence. In the NAC sequence of α -synuclein and in AChE_{586–599}, the helical propensity predicted by both PHD and the present method (*P*[β /low]) was relatively high compared to non-amyloidogenic counterparts. This intriguing outcome concurs with the view that the amyloid fibril forming propensity of a given protein sequence is related to its ambivalent nature with respect to adopting a α -helix or β -strand conformation.

In the native state, myoglobin is a highly soluble globular protein with a secondary structure that is almost exclusively α -helical punctuated by short loops that link the helices. Its entire sequence is devoid of the β -strands associated with amyloid fibril formation, and no β -strand propensities can be detected in this protein by the PHD algorithm (Fig. 2A).

Nevertheless, Fändrich et al. (2001) have demonstrated that, under partially destabilizing in vitro conditions, this protein can be induced to form β -stranded fibrils that are virtually identical to those seen in disease-associated amyloid fibrils. Given the profound implications of this experimental finding, we pondered whether our TC-based H β P method would detect any significant H β P in the myoglobin sequence. Indeed, both EVLIRLF and TVVLTAL were predicted by the H β P method to show the strongest nonnative β -strand propensity in high TC environments (Fig. 2B) notwithstanding the fact that these sequences are α -helical in the native state. Other helical sequences (VLNVWGKVEA, IKYLEFIS, and IIHVLHKS) also revealed significant H β P. Fändrich et al. (2003) recently reported an independent peptide fragment containing IKYLEFIS and IIHVLHKS as amyloidogenic, although it is known as a stable helical element in the protein and lacks clear polar-hydrophobic sequence pattern. In accordance with these experimental findings, the present computational analysis shows remarkable conformational ambivalence (helix for low and β -strand for high TCs) for this region. These previously undetected H β P may explain why such a protein that is predominantly helical in the native state is still capable of forming fibrillar aggregates.

The notion that a small number of compact sequences, such as EVLIRLF and TVVLTAL in myoglobin, can provide sufficient driving force for amyloid fibril formation might seem implausible. These two seven-residue sequences represent a small fraction of horse myoglobin's 153 residues, and they are well separated in terms of sequence as well as spatially in the crystalline state. On the other hand, the finding (Tjernberg et al. 1996) that the highly amyloidogenic behavior of A β can be arrested by replacing its KLVFF pentapeptide with the helix stabilizing AAVFA attests to the apparent delicate balance between normal and amyloid structures. Likewise, our H β P method detected high β -strand propensities in only a few regions of A β (Fig. 3A). This invites speculation whether removal of EVLIRLF and TVVLTAL in myoglobin, either by replacing them with helix-promoting residues (e.g., alanine) or by deleting them altogether, would attenuate the tendency of myoglobin for fibril formation under the same experimental conditions used by Fändrich et al. (2001). Alternatively, addition of EVLIRLF and TVVLTAL peptides to these myoglobin solutions might serve as " β -sheet breakers," much like AcQKLVFFNH₂ was shown by Tjernberg et al. (1996) to halt polymerization of A β and subsequent formation of fibrils.

H β P in globular domains

Despite the growing number of proteins shown in various experiments to be amyloidogenic, no definitive patterns have been found as to sequence specificity or to the thresh-

Table 4. Secondary structure propensity of known amyloidogenic and nonamyloidogenic subsequences

Protein	Subsequence	Amyloid formation	Reference	HβP method		PHD	
				P(α/low)	P(β/high)	prH	prE
β-amyloid	KL VFF	Strong	Tjernberg et al. 1996	0.6	1.0	0.0	0.8
	AAV F A	Weak		0.7	0.7	0.1	0.6
IAPP	NFL VH	Yes	Mazor et al. 2002	0.7	0.8	0.8	0.0
	FL VHS	Yes		0.5	0.7	0.7	0.0
NAC	QKTVEGAGSIAAATGEV	No	El-Agnaf et al. 1998;	0.4	0.2	0.1	0.6
sequence of α-synuclein	(19–35)						
	VT NVGGAVVTG V TAVA	Yes	Bodles et al. 2001	0.4	0.6	0.4	0.2
	(3–18)						
	VT GTAVA Q KTV	Yes	Giasson et al. 2001	0.6	0.7	0.4	0.3
	(11–22)						
	GA VVTG V TAVA	Yes		0.5	0.7	0.4	0.3
	(8–18)						
	Core, VTGVTAVA ^a	?		0.6	0.8	0.4	0.3
	(11–18)						
AchE _{586–599}	AE FHRWSS Y MV H WK	Yes	Cottingham et al. 2002;	0.8	0.7	0.7	0.1
BuChE _{573–596}	AGFHRWNNY M MDWK	No	Cottingham et al. 2003	0.5	0.2	0.8	0.0

^a This sequence represents the overlapping core of three amyloidogenic fragments, for which the exact core fragment was not identified.

Calculated HβP and PHD propensities (residue averages) listed below are shown in graphical format in Fig. 3. For the purpose of comparison to HβP method, prH and prE of PHD prediction were normalized to a zero-to-one scale. Boldface letters represent amyloidogenic fragments and their β propensity.

old level of additional β-strands that can trigger amyloid fibril formation. Analysis of 2358 nonhomologous domains in the SCOP20 database revealed that the *domain* HβP for the majority of domains is in the 0.2 to 0.5 range (Fig. 5A), meaning that 20% to 50% of residues were predicted by the present method to possess significant β-strand propensity ($P[\beta|high] > 0.5$) even though they are not β-strand in the native state. Because the present method is designed to detect nonnative as opposed to native β-strand propensity, the HβP will be greater for helical domains than for β-rich

domains (Fig. 5A). For example, the predicted *domain* HβP was low for β-rich proteins (viz., SH3 domain and Immunoglobulin-light chain) and noticeably higher for Aβ and other helical proteins (viz., insulin, myoglobin; Table 5).

A χ^2 analysis comparing the *domain* HβP among 23 proteins shown experimentally as amyloidogenic and the 2358 SCOP20 domains indicated similar distributions with no significant difference between the two distributions (Fig. 5, including statistical parameters). The lowest level of *domain* HβP in known amyloidogenic proteins is 0.2, whereas most globular domains in SCOP20 are predicted to possess *domain* HβP > 0.2. The average *domain* HβP was identical ($H\beta P_{avg} = 0.30$) for both the SCOP20 domains and the 23 known amyloidogenic proteins (Fig. 5A,B). Although other factors are likely to affect fibril formation, the finding that significant HβP exists in most SCOP20 domains corroborates the notion that amyloid fibril formation is a generic feature of proteins (Chiti et al. 1999). Normally benign proteins become toxic when they undergo fibrilization (Bucciantini et al. 2002). To the extent that in vitro observations reflect in vivo behavior (Couzin 2002), the results summarized in Figure 5 suggest that amyloid fibrils can be induced in most if not all proteins during the course of their biological function through tertiary structural rearrangement.

The high prevalence of nonnative β-strand propensity in local amino-acid sequences of proteins (Fig. 5) is fascinating yet disturbing when seen from the perspective that fibrillar aggregates (or their intermediates) are likely toxic in humans. Structural plasticity in a local sequence has been observed, even by a single-site mutation (Cordes et al.

Figure 4. Comparison of residue β propensity to the distribution of binding motifs in the hIAPP sequence. Binding data of 28 consecutive overlapping decamers were retrieved from Fig. 3b in Mazor et al. (2002). Predicted β propensity was obtained from $P(\beta|high)$ of the present HβP method and from prE of the PHD method. Original prE from PHD (see Fig. 3C) was normalized onto a zero-to-one scale for comparison with HβP. Values of $P(\beta|high)$ and prE for the central residue in each decamer were compared with the binding data corresponding to that decamer (Table 3).

Figure 5. The domain H β P of globular proteins and known amyloidogenic proteins. (A) The 2358 SCOP20 domains were analyzed. Red bars represent the distribution of all α domains. Blue is β , and white represents the mixture of α and β (α/β , $\alpha + \beta$, and multidomains) domains. The average H β P of 2358 domains is 0.30. (B) H β P of 23 proteins shown experimentally as amyloidogenic. PDB ID and protein name are listed in Table 5. The average H β P is also 0.30. Comparison was made at 12 different H β P levels ranging from 0.0 to 0.6. χ^2 distribution of domain H β P between A and B is 12.6, degrees of freedom are $(12 - 1)(2 - 1) = 11$, and $P = 0.31 > 0.05$.

2000), thus suggesting that new protein folds can evolve from existing folds without drastic or large-scale mutagenesis. A plausible interpretation is that amyloid fibrils are by-products resulting from the inherent structural plasticity of local amino-acid sequences. The ambivalent nature of local sequences as to their structural propensities might imply that peptide sequences represent a neutral pool for protein evolution that depends on the appropriate control system (e.g., molecular chaperones) to suppress protein misfolding and aggregation. In concert with the control system, the H β P of a local sequence might represent another driving force in protein evolution. This premise is supported by the significant degree of domain H β P detected in most SCOP20 sequences (Fig. 5). Identification of the H β P of local sequences provides insight into the role of amyloid fibrils in protein evolution and should contribute toward progress in our battle against amyloid diseases and related conditions.

Therapeutic implications

Although several therapeutic targets (e.g., the secretases) have been identified to block the amyloid cascade upstream of fibrillar formation (Wolfe 2002), no clinically effective drugs for these targets have yet appeared. Somewhat counter-intuitively, small peptides composed of these same local sequences associated with strong H β P have been shown by in vitro experiments (Soto et al. 1998; Citron 2002; Findeis 2002) to block amyloid fibril formation in full-length proteins. These peptides, aptly called β -sheet breakers, are based on A β residues 17–21, which constitute the central hydrophobic core that is believed essential for A β assembly (Tjernberg et al. 1996; Soto et al. 1998). Encouragingly, recent in vivo studies (Permanne et al. 2002) using two different transgenic mouse models have demonstrated that systematic administration of a pentapeptide β -sheet breaker can reduce amyloid load and cerebral damage in Alzheimer's disease. These small peptides exhibited good brain penetration, reduced A β deposition, increased neuronal survival, and decreased brain inflammation associated with amyloid deposition.

Although promising, the β -sheet breaker approach requires prior knowledge of those sequences associated with H β P. Ready access to this information has been hampered by the absence of a consensus sequence among disease-associated amyloid forming proteins, confounded by difficulties encountered in determining fibril structures at the

Table 5. The domain H β P of 23 known amyloidogenic proteins

PDB ID	Protein name	Domain H β P
1ba6	β -amyloid	0.50
1mhi	Insulin	0.49
1aye	Ada2h	0.41
1ymb	Myoglobin	0.40
1n9d	Prolactin	0.38
1qlx	Human prion protein	0.35
1hiv	Gelsolin	0.33
1ggt	Coagulation factor XIII	0.32
1tca	Lipase B	0.32
1bzb	Calcitonin	0.32
3pte	Transpeptidase	0.29
2ach	Antichymotrypsin	0.29
1n76	Lactotransferrin	0.28
1lhl	Lysozyme	0.28
1jc9	Fibrinogen	0.26
2try	Transthyretin	0.25
1ayg	Cytochrome c552	0.23
1av1	Apolipoprotein	0.23
1oez	Superoxide dismutase	0.22
1pnj	SH3 domain	0.21
1g96	Cystatin C	0.21
12e8	Immunoglobulin-light chain	0.18
1im9	β_2 -microglobulin	0.17

The native secondary structure was determined using PDB structures. The average value of domain H β P is 0.30.

molecular level using experimental techniques. These obstacles make it extremely difficult to identify target sites of β -sheet breakers and to design effective β -sheet breakers. It is hoped that the present H β P method for detecting H β P will offer some guidance in addressing this problem.

Future directions

The present method demonstrated an exceptional sensitivity to detect H β P in local regions of protein sequences. Our analysis of H β P in globular domains revealed a high degree of association with known amyloidogenic peptides. This new approach enables us to carry out proteome-wide analysis of amyloidogenic propensity in protein sequence. Currently, we are developing a Web-accessible interface that implements our TC-based H β P method within an artificial neural network (ANN) to enable fast and accurate prediction of H β P for any protein or peptide sequence along a continuum of variable TC values. This Web site will also feature a knowledge base that contains a database of short sequences that are predicted by our ANN-based H β P tool to possess strong H β P. This database, which will be updated as more protein structural information becomes available, is intended to offer guidance toward the discovery of therapeutic agents that inhibit β -strand formation and aggregation.

Materials and methods

Nonhomologous heptapeptide library

A total of 453,787 amino-acid sequence fragments, each seven residues in length, were extracted from 2589 domains of SCOP20, version 1.57 (Brenner et al. 2000) after exclusion of membrane proteins, small proteins, and proteins with incomplete structural information. SCOP20, a collection of protein domains that exhibit <20% sequence identity between any two members, provided a rich source of nonhomologous sequence contexts from diverse tertiary environments. These SCOP20 fragments were used as the template pool to calculate the TC-dependent secondary structure propensity of a query sequence. The number of TCs for each fragment was calculated from the experimental structure of the corresponding protein retrieved from the Protein Data Bank (PDB; <http://www.rcsb.org/pdb/>). A TC is defined as two heavy (non-H) atoms ≤ 4 Å apart and separated by more than four residues in sequence (Fischer and Marqusee 2000). The TCs of the middle five residues within each seven-residue sequence were counted and then sorted categorically into bins designated *low*, *intermediate*, and *high*. The *low/intermediate* and *intermediate/high* bound-

aries were defined respectively as $TC_{avg} - 20\%$ and $TC_{avg} + 20\%$, where TC_{avg} is the average number of TCs. Because individual amino acids will differ with respect to side-chain length, composition, and hydrophobicity, the TC_{avg} value associated with each of the 20 common amino acids was precalculated from the original 453,787 fragments in the SCOP20 database (Table 6). The values of TC_{avg} for these amino acids are seen to vary in a consistent manner, namely, highest for those with aromatic side chains (W, Y, F) and lowest for those with small side chains (A, G). The value of TC_{avg} for a given seven-residue fragment is calculated as the sum of average TCs of the constituent amino acids. The TC of a query sequence is classified as high or low by comparing TC_{avg} and the actual TC of the query for the middle five residues. The TC_{avg} of the middle five residues in a query sequence is calculated from data in Table 6. The secondary structure of the central residue in each seven-residue fragment was calculated using the Database of Secondary Structure in Proteins (DSSP) program (Kabsch and Sander 1983).

Calculating $P(\alpha|low)$ and $P(\beta|high)$

For a given seven-residue query sequence, the 30 top-scoring templates containing a central residue identical to the query were selected from 453,787 fragments in SCOP20 (Fig. 1). Preliminary tests conducted by using a variable number (20 to 60) of templates found 30 to strike a balance between prediction accuracy and computational efficiency. The PAM30 (Schwartz and Dayhoff 1978) substitution matrix was used for gapless alignment between a query and the templates. In cases in which <30 templates were found that achieved a minimum PAM30 score of 15, only those templates above this cutoff were retained for further analysis.

After sorting the maximum 30 templates into three separate bins designated *low*, *intermediate*, and *high* in terms of number of TCs as described above, the secondary structure of the center residue of each template in the separate bins was analyzed. The occurrences of α -helix in low TC bins, $P(\alpha|low)_{temp}$, and β -strand in high TC bins, $P(\beta|high)_{temp}$, were calculated from the templates in low and high TC bins, respectively (Fig. 1). The templates in the *intermediate* bin were discarded. To predict $P(\alpha|low)$ and $P(\beta|high)$ of a query from the corresponding information obtained from the templates, we developed statistical models by defining two additional variables, $N(low)_{temp}$ and $N(high)_{temp}$, that correspond to the total number of templates in low and high TCs, respectively (Fig. 6A,B). The inclusion of $N(low)_{temp}$ and $N(high)_{temp}$ improves our ability to predict $P(\alpha|low)_{query}$ from $P(\alpha|low)_{temp}$ and $P(\alpha|high)_{query}$ from $P(\alpha|high)_{temp}$.

Removing each of the 453,787 SCOP20 fragments one at a time as a test query, we searched the remaining SCOP20 453,786 fragments to retrieve the templates. Among the 453,787 SCOP20 fragments, we found 175,503 test queries for which $N(low)_{temp} \geq 3$ and $P(\alpha|low)_{temp} > 0.5$. We proceeded to plot the occurrence of helix in the queries, $P(\alpha|low)_{query}$, against their $N(low)_{temp}$ separately for three different ranges of $P(\alpha|low)_{temp}$ (Fig. 6A). For example, the value indicated by an arrow in Figure 6A was determined from 782 test query fragments with low TCs. They repre-

Table 6. Average TC number for each of the 20 common amino acids, based on analysis of 453,787 amino acids in SCOP20

Amino acid	L	S	A	Q	E	K	D	N	R	I	F	T	M	P	G	W	C	V	Y	H
TC_{avg}	9.9	8.3	6.4	10.2	8.6	8.0	9.1	10.6	15.9	11.0	18.5	9.3	11.9	7.0	6.5	25.9	12.7	10.1	21.7	16.2

See the Materials and Methods section for the definition of TC and TC_{avg} .

Figure 6. Correlations for $P(\alpha|low)$ and $P(\beta|high)$ between queries and their templates. (A) Values of $P(\alpha|low)_{query}$ are plotted against $N(low)_{templates}$ at discrete levels of $P(\alpha|low)_{templates}$. A total of 175,503 SCOP20 fragments found in low TCs are used as test queries. For explanation of point designated by arrow, see Materials and Methods. The curves were fitted to the following equations: $y = 0.076\ln(x) + 0.77$ (circles), $y = 0.13\ln(x) + 0.58$ (triangles), and $y = 0.15\ln(x) + 0.38$ (squares). (B) Values of $P(\beta|high)_{query}$ are plotted against $N(high)_{templates}$ at discrete levels of $P(\beta|high)_{templates}$. A total of 113,680 SCOP20 fragments found in high TCs were used as test queries. The curves were fitted to the following equations: $y = 0.17\ln(x) + 0.54$ (circles), $y = 0.19\ln(x) + 0.42$ (triangles), and $y = 0.16\ln(x) + 0.34$ (squares).

sent the collection of test queries that have identical levels of $P(\alpha|low)_{temp}$ and $N(low)_{temp}$, namely, $P(\alpha|low)_{temp}$ in the 0.5 to 0.7 range and $N(low)_{temp} = 6$. Because 566 of these 782 test queries are α -helical in the native state, $P(\alpha|low)_{query} = 0.72$ (i.e., 566/782). From the 453,787 SCOP20 fragments, we found 113,680 test queries for which $N(high)_{temp} \geq 3$ and $P(\alpha|high)_{temp} > 0.5$. Similarly, values of $P(\alpha|high)_{query}$ versus $N(high)_{temp}$ were plotted separately for three different ranges of $P(\alpha|high)_{temp}$ (Fig. 6B).

The curves so obtained (Fig. 6A,B) were each fitted to a non-linear regression equation. The values of $P(\alpha|low)_{query}$ and $P(\alpha|high)_{query}$ for each query sequence, such as the sequence GEAVELA shown in Figure 1, were determined from these equations (Fig. 6A,B, respectively). Inspection of the curves in Figure 6 reveals, as expected, that the statistical quality of the predicted propensity of a query/test sequence improves as the number of templates, $N(low)_{temp}$, or $N(high)_{temp}$, increases and as the propensity of the corresponding templates, $P(\alpha|low)_{temp}$ or $P(\alpha|high)_{temp}$, approaches unity.

Values of $P(\alpha|low)$ and $P(\alpha|high)$ for a full-length protein query sequence were calculated by using a sliding seven-residue win-

dow. The first three residues at each terminus of the protein were excluded from the prediction process because they lack the minimum three residues on one side needed to assign sequence context. Predictions of secondary structure as calculated by PHD algorithms were carried out by accessing the Predict Protein Server (http://www.embl-heidelberg.de/predictprotein/submit_def.html). The domain $H\beta P$ parameter was calculated as

$$\frac{N\{non\beta, P(\beta|high) > 0.5\}}{N(non\beta)}$$

where $N(non\beta)$ refers to the number of residues in a given domain with secondary structure that is not β -strand in the native structure and $N\{non\beta, P(\beta|high) > 0.5\}$ refers to the subset of these for which $P(\alpha|high) > 0.5$.

Acknowledgments

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

- Balbirnie, M., Grothe, R., and Eisenberg, D.S. 2001. An amyloid-forming peptide from the yeast prion Sup35 reveals a dehydrated β -sheet structure for amyloid. *Proc. Natl. Acad. Sci.* **98**: 2375–2380.
- Benyamini, H., Gunasekaran, K., Wolfson, H., and Nussinov, R. 2003a. β_2 -Microglobulin amyloidosis: Insights from conservation analysis and fibril modeling by protein docking techniques. *J. Mol. Biol.* **330**: 159–174.
- . 2003b. Conservation and amyloid formation: A study of the gelsolin-like family. *Proteins* **51**: 266–282.
- Berezovsky, I.N. and Trifonov, E.N. 2001. Van der Waals locks: Loop-n-lock structure of globular proteins. *J. Mol. Biol.* **307**: 1419–1426.
- Bodles, A.M., Guthrie, D.J., Greig, B., and Irvine, G.B. 2001. Identification of the region of non- $\alpha\beta$ component (NAC) of Alzheimer's disease amyloid responsible for its aggregation and toxicity. *J. Neurochem.* **78**: 334–395.
- Brenner, S.E., Koehl, P., and Levitt, M. 2000. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.* **28**: 254–256.
- Bucciantini, M., Giannoni, E., Chiti, F., Baroni, F., Formigli, L., Zurdo, J., Taddei, N., Ramponi, G., Dobson, C.M., and Stefani, M. 2002. Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature* **416**: 507–511.
- Chiti, F., Webster, P., Taddei, N., Clark, A., Stefani, M., Ramponi, G., and Dobson, C.M. 1999. Designing conditions for in vitro formation of amyloid protofibrils and fibrils. *Proc. Natl. Acad. Sci.* **96**: 3590–3594.
- Chiti, F., Stefani, M., Taddei, N., Ramponi, G., and Dobson, C.M. 2003. Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature* **424**: 805–808.
- Citron, M. 2002. Alzheimer's disease: Treatments in discovery and development. *Nat. Neurosci.* **5**: 1055–1057.
- Cordes, M.H., Burton, R.E., Walsh, N.P., McKnight, C.J., and Sauer, R.T. 2000. An evolutionary bridge to a new protein fold. *Nat. Struct. Biol.* **7**: 1129–1132.
- Cottingham, M.G., Hollinshead, M.S., and Vaux, D.J. 2002. Amyloid fibril formation by a synthetic peptide from a region of human acetylcholinesterase that is homologous to the Alzheimer's amyloid- β peptide. *Biochemistry* **41**: 13539–13547.
- Cottingham, M.G., Voskuil, J.L., and Vaux, D.J. 2003. The intact human acetylcholinesterase C-terminal oligomerization domain is α -helical in situ and in isolation, but a shorter fragment forms β -sheet-rich amyloid fibrils and protofibrillar oligomers. *Biochemistry* **42**: 10863–10873.
- Couzin, J. 2002. Harmless proteins twist into troublemakers. *Science* **296**: 28–29.
- El-Agnaf, O.M., Jakes, R., Curran, M.D., Middleton, D., Ingenito, R., Bianchi, E., Pesse, A., Neill, D., and Wallace, A. 1998. Aggregates from mutant and wild-type α -synuclein proteins and NAC peptide induce apoptotic cell death

- in human neuroblastoma cells by formation of β -sheet and amyloid-like filaments. *FEBS Lett.* **440**: 71–75.
- Fändrich, M., Fletcher, M.A., and Dobson, C.M. 2001. Amyloid fibrils from muscle myoglobin. *Nature* **410**: 165–166.
- Fändrich, M., Forge, V., Buder, K., Kittler, M., Dobson, C.M., and Diekmann, S. 2003. Myoglobin forms amyloid fibrils by association of unfolded polypeptide segments. *Proc. Natl. Acad. Sci.* **100**: 15463–15468.
- Findeis, M.A. 2002. Peptide inhibitors of β -amyloid aggregation. *Curr. Top. Med. Chem.* **2**: 417–423.
- Fischer, K.F. and Marqusee, S. 2000. A rapid test for identification of autonomous folding units in proteins. *J. Mol. Biol.* **302**: 701–712.
- Giasson, B.I., Murray, I.V., Trojanowski, J.Q., and Lee, V.M. 2001. A hydrophobic stretch of 12 amino acid residues in the middle of α -synuclein is essential for filament assembly. *J. Biol. Chem.* **276**: 2380–2386.
- Hoppener, J.W., Ahren, B., and Lips, C.J.M. 2000. Islet amyloid and type 2 diabetes mellitus. *N. Engl. J. Med.* **343**: 411–419.
- Jimenez, J.L., Guijarro, J.I., Orlova, E., Zurdo, J., Dobson, C.M., Sunde, M., and Saibil, H.R. 1999. Cryo-electron microscopy structure of an SH3 amyloid fibril and model of the molecular packing. *EMBO J.* **18**: 815–821.
- Kabsch, W. and Sander, C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**: 2577–2637.
- Kallberg, Y., Gustafsson, M., Persson, B., Thyberg, J., and Johansson, J. 2001. Prediction of amyloid fibril-forming proteins. *J. Biol. Chem.* **276**: 12945–12950.
- Mangione, P., Sunde, M., Giorgetti, S., Stoppini, M., Esposito, G., Gianelli, L., Obici, L., Asti, L., Andreola, A., Viglino, P., et al. 2001. Amyloid fibrils derived from the apolipoprotein A1 Leu174Ser variant contain elements of ordered helical structure. *Protein Sci.* **10**: 187–199.
- Mazor, Y., Gilead, S., Benhar, I., and Gazit, E. 2002. Identification and characterization of a novel molecular-recognition and self-assembly domain within the islet amyloid polypeptide. *J. Mol. Biol.* **322**: 1013–1024.
- Minor Jr., D.L. and Kim, P.S. 1996. Context-dependent secondary structure formation of a designed protein sequence. *Nature* **380**: 730–734.
- Pan, X.-M., Niu, W.-D., and Wang, Z.-X. 1999. What is the minimum number of residues to determine the secondary structural state? *J. Protein Chem.* **18**: 579–584.
- Paz, M. and Serrano, L. 2004. Sequence determinants of amyloid fibril formation. *Proc. Natl. Acad. Sci.* **101**: 87–92.
- Permanne, B., Adessi, C., Saborio, G.P., Fraga, S., Frossard, M.-J., Dorpe, J.V., Dewachter, I., Banks, W.A., Leuven, F.V., and Soto, C. 2002. Reduction of amyloid load and cerebral damage in a transgenic mouse model of Alzheimer's disease by treatment with a β -sheet breaker peptide. *FASEB J.* **8**: 860–862.
- Plaxco, K.W., Simons, K.T., and Baker, D. 1998. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**: 985–994.
- Rost, B. 1996. PHD: Predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol.* **266**: 525–539.
- Sacchettini, J.C. and Kelly, J.W. 2002. Therapeutic strategies for human amyloid diseases. *Nat. Rev. Drug Discovery* **1**: 267–275.
- Schwartz, R.M. and Dayhoff, M.O. 1978. Matrices for detecting distant relationships. In *Atlas of protein sequence and structure* (ed. M.O. Dayhoff), pp. 353–358. National Biomedical Research Foundation, Washington, DC.
- Soto, C., Sigurdsson, E.M., Morelli, L., Kumar, R.A., Castano, E.M., and Frangione, B. 1998. β -Sheet breaker peptides inhibit fibrillogenesis in a rat brain model of amyloidosis: Implications for Alzheimer's therapy. *Nat. Med.* **4**: 822–826.
- Spillantini, M.G., Schmidt, M.L., Lee, V.M., Trojanowski, J.Q., Jakes, R., and Goedert, M. 1997. α -Synuclein in Lewy bodies. *Nature* **388**: 839–840.
- Spillantini, M.G., Crowther, R.A., Jakes, R., Hasegawa, M., and Goedert, M. 1998. α -Synuclein in filamentous inclusions of Lewy bodies from Parkinson's disease and dementia with lewy bodies. *Proc. Natl. Acad. Sci.* **95**: 6469–6473.
- Tjernberg, L.O., Naslund, J., Lindqvist, F., Johansson, J., Karlstrom, A.R., Thyberg, J., Terenius, L., and Nordstedt, C. 1996. Arrest of β -amyloid fibril formation by a pentapeptide ligand. *J. Biol. Chem.* **271**: 8545–8548.
- Ueda, K., Fukushima, H., Masliah, E., Xia, Y., Iwai, A., Yoshimoto, M., Otero, D.A., Kondo, J., Ihara, Y., and Saitoh, T. 1993. Molecular cloning of cDNA encoding an unrecognized component of amyloid in Alzheimer disease. *Proc. Natl. Acad. Sci.* **90**: 11282–11286.
- Wolfe, M.S. 2002. Therapeutic strategies for Alzheimer's disease. *Nat. Rev. Drug Discovery* **1**: 859–866.
- Zhou, X., Alber, F.A., Folkers, G., Gonnet, G.H., and Chelvanayagam, G. 2000. An analysis of the helix-to-strand transition between peptides with identical sequence. *Proteins* **41**: 248–256.