BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 17 2007, pages 2218–2225
doi:10.1093/bioinformatics/btm325

Structural bioinformatics

Identification of amyloid fibril-forming segments based on structure and residue-based statistical potential

Zhuqing Zhang[1,2], Hao Chen[1,2] and Luhua Lai[1,2,*]

[1]Beijing National Laboratory for Molecular Sciences, State Key Laboratory for Structural Chemistry of Unstable and Stable Species, College of Chemistry and Molecular Engineering and [2]Center for Theoretical Biology, Peking University, Beijing 100871, China

Received on December 6, 2006; revised on May 29, 2007; accepted on June 15, 2007

Advance Access publication June 28, 2007

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: Experimental evidence suggests that certain short protein segments have stronger amyloidogenic propensities than others. Identification of the fibril-forming segments of proteins is crucial for understanding diseases associated with protein misfolding and for finding favorable targets for therapeutic strategies.

Result: In this study, we used the microcrystal structure of the NNQQNY peptide from yeast prion protein and residue-based statistical potentials to establish an algorithm to identify the amyloid fibril-forming segment of proteins. Using the same sets of sequences, a comparable prediction performance was obtained from this study to that from 3D profile method based on the physical atomic-level potential ROSETTADESIGN. The predicted results are consistent with experiments for several representative proteins associated with amyloidosis, and also agree with the idea that peptides that can form fibrils may have strong sequence signatures. Application of the residue-based statistical potentials is computationally more efficient than using atomic-level potentials and can be applied in whole proteome analysis to investigate the evolutionary pressure effect or forecast other latent diseases related to amyloid deposits.

Availability: The fibril prediction program is available at ftp://mdl.ipc.pku.edu.cn/pub/software/pre-amyl/

Contact: lhlai@pku.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Amyloid fibrillar aggregates, caused by the misfolding of peptides and proteins, are associated with a range of human disorders such as Alzheimer's disease, Parkinson's disease, transmissible spongiform encephalopathies and type II diabetes (Dobson, 1999; Rochet and Lansbury, 2000; Ross and Poirier, 2004). Likewise, some non-pathogenic peptides or proteins (Dobson, 2002; Fandrich *et al*., 2001; Guijarro *et al*., 1998) and designed peptides (Lopez De La Paz *et al*., 2002) can also form amyloid fibrils under appropriate conditions. Although the amyloidogenic proteins do not share any sequence homology or common native fold patterns, they are remarkably similar in their cross β structures (Dobson, 1999; Rochet and Lansbury, 2000), with β-sheet backbones perpendicular to and hydrogen bonds parallel to the fibril axis, and can bind Congo red with characteristic birefringence (Klunk *et al*., 1999). This suggests that the underlying mechanism of fibril assembly may be common to all amyloidogenic proteins. However, due to the non-crystalline and insoluble nature of amyloid fibrils, and to their complex formation processes, a detailed understanding of the fibrils formation remains limited.

Some studies have implied that assembly into amyloid-like fibrils is an inherent property of polypeptides, irrespective of their sequence (Dobson, 1999, 2001). However, it is obvious that some sequences are much more amyloidogenic than others. Moreover, some short peptides possess the same amyloid properties as full length proteins (Balbirnie *et al*., 2001; Tenidis *et al*., 2000), and some very short specific stretches have been considered to be the regions responsible for aggregation, as they can change the amyloidogenic propensities of polypeptides by facilitating or inhibiting fibril formation (Ivanova *et al*., 2004; Ventura *et al*., 2004). These data suggest that peptide sequence can influence amyloid fibril formation, and has inspired the recent development of a number of algorithms and models that predict the amyloidogenic or aggregation propensities of polypeptides or proteins (Bemporad *et al*., 2006; Caflisch, 2006). For example, in mutagenesis experiments where individual residues of amyloidogenic peptide STVIIE were systematically replaced with all natural amino acids except cysteine, Lopez de la Paz and Serrano (Lopez de la Paz and Serrano, 2004) derived a sequence pattern to identify the amyloidogenic potentiality of six-residue segments. Likewise, the statistical mechanics algorithm TANGO, developed by Serrano *et al*. (Fernandez-Escamilla *et al*., 2004), was based on secondary structure propensity and estimation of desolvation penalty to predict β-aggregating regions of a protein sequence and mutational effects, and has recently been used to investigate the aggregation propensity of proteins in 28 complete proteomes spanning all kingdoms of life (Rousseau *et al*., 2006). Based on a linearly combined function of hydrophobicity, α-helical propensity, β-sheet propensity and charge effect, Pawar *et al*. (Pawar *et al*., 2005) proposed a

[*]To whom correspondence should be addressed.

property-based method for assaying 'aggregation-prone' segments of proteins. Yoon and Welsh (Yoon and Welsh, 2004) developed an algorithm to detect the hidden β propensity (HβP) in terms of its tertiary structure context. In addition, De Groot and co-workers (De Groot *et al*., 2005) developed an approach to identify 'hot-spots' of aggregation in globular and unstructured disease-linked polypeptides by using the empirical scale of aggregation propensity for all 20 natural residues, which was determined through a complete set of mutations of Phe 19 in the central hydrophobic cluster (CHC) of Aβ. Saiki *et al*. (Saiki *et al*., 2006) built an interaction-based structure model and Galzitskaya and coworkers (Galzitskaya *et al*., 2006) described another algorithm based on the stacking extent of residues to detect amyloidogenic regions in proteins. All of these methods rely on diverse assumptions, and have been applied successfully to some extent.

Of primary interest to our study was the recently developed 3D profile method (Thompson *et al*., 2006) that identifies fibril-forming segments based on the microcrystal structure of the fibril-forming peptide NNQQNY (Nelson *et al*., 2005). In the present study, instead of the physical-based all atom potential used by Thompson *et al*., we adopted residue-based statistical potential to predict the fibril-forming propensity of different hexapeptides by threading them onto the microcrystal structure of the NNQQNY peptide. Residue-based statistical potentials, which have been successfully used in folding recognition (Bowie *et al*., 1991; Jones *et al*., 1992) and *ab initio* structure prediction (Skolnick *et al*., 1997), are knowledge-based potentials extracted residue interactions directly from known protein structures. We have tested several residue-based potentials and obtained prediction results comparable to those of the physical-based atomic level potential ROSETTADESIGN (Kuhlman and Baker, 2000). The atomic level physical potential is limited by available computational power when encountered with large calculation systems as it involves multiple classical energy terms, each of which includes many empirical parameters and functions. However, residue-based statistical potential is simple and suitable for large-scale calculations such as a proteome analyses, as in this study we analyzed the entire yeast proteome to predict amyloidgenic peptide segments using residue-based statistical calculations.

2 METHODS

2.1 Template structures

We constructed the template library consisting of 2511 structures with a slight perturbation in coordinates of the microcrystal structure of the NNQQNY peptide (PDB code 1YJO) (Nelson *et al*., 2005) to compare the prediction performance in this study with those obtained using ROSETTADESIGN as the scoring function (Thompson *et al*., 2006). Each template structure consisted of two β-sheets, one with 9 strands and the other with 12 strands, where the basic topology of intra-sheet strands stacked in parallel with two anti-parallel β-sheets was retained. The perturbation of the templates was made by shifting the two β-sheets according to three orthogonal directions *x*, *y* and *z* with certain interval, as shown in Figure 1. The shift is 7.5 Å along the *x* axis with the interval of 0.25 and 4.5 Å along the *y* axis with the interval of 0.25 and 2.4 Å along the *z* axis with the interval 1.2 Å. Thus in all the template structures, the central strand of the nine-strand sheet was consistently buried in the fibril environment (with a 15 Å cut-off for residue-base statistical potential energy calculation) and its interactions with other segments were calculated in energy scoring.

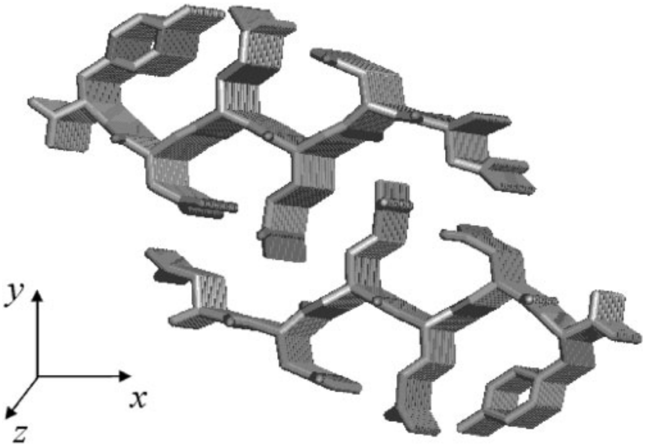

Fig. 1. Schematic representation of one template structure. The two β-sheets are oriented in antiparallel, one includes 9 strands, and the other includes 12 strands which stack in parallel within each sheet. The *x*, *y* and *z* axes denote the direction of the shift between the two β-sheets for the construction of template structures.

2.2 Sequence data sets

The data sets used for examining prediction performance basically came from those described in Thompson *et al*. work (Thompson *et al*., 2006). These data sets included AmylHex consisting of 67 six-residue peptides known to form fibrils and 91 six-residue peptides that cannot. The data sets also included AmylFrag containing 45 amyloidogenic fragments of proteins identified in various experiments, from which 449 unique hexameric peptides were collected by scanning the fragments with a six-residue window. The control set of sequences contained 670 unique hexapeptides and were obtained from a set of full-length proteins (myoglobin with PDB code 2BLI, Ribonuclease A with PDB code 1C9X, Human prion protein with PDB code 1HJM, insuline with PDB code 1ZEH, β_2-microglobulin with PDB code 1LDS, α-synuclein with PDB code 1XD8 and Human Transthyretin with PDB code 1FHN) by scanning with a six-residue window and excluding those in AmylHex or AmylFrag sets. The ability of the control set sequences to form fibrils is unknown.

2.3 Interaction energy calculations

Each expected six-residue peptide was mapped onto each of the template structures. The residue-based statistical potential was used to evaluate the interaction energy score of the central strand in the nine-strand β-sheet with other strands. The lowest energy score obtained from the template structures was then used to assay the fibril-forming propensity of this peptide.

Residue-based statistical potential is the statistical mean force exacted from experimentally solved protein structures (Tanaka and Scheraga, 1976). A distance-dependent, pair-wise statistical potential is generally expressed as

$$u(i,j,r) = -RT\ln\frac{N_{\text{obs}}(i,j,r)}{N_{\exp}(i,j,r)} \quad (1)$$

where R is the gas constant, T is the temperature, $N_{\text{obs}}(i,\ j,\ r)$ and $N_{\text{exp}}(i,\ j,\ r)$ are the observed and expected numbers of residues or atoms pair $(i,\ j)$ contacts in distance shell $r–\Delta r$. In this study, C_β was used to stand for the residue in statistics. The difference in various statistical potentials primarily comes from how to calculate $N_{\text{exp}}(i,\ j,\ r)$ from a number of structures. We first tested three common statistical potentials, KBP (Lu and Skolnick, 2001), RAPDF (Samudrala and Moult, 1998) and DFIRE (Zhou and Zhou, 2002) to calculate energy. Their $N_{\text{exp}}(i,\ j,\ r)$ were obtained from quasi-chemical approximation, statistical average and the physical state of finite ideal-gas, respectively.

KBP:

$$N_{\text{exp}}(i,j,r) = x_i x_j N_{\text{obs}}(r) \tag{2}$$

where $N_{\text{obs}}(r) \equiv \sum_{i,j} N_{\text{obs}}(i,j,r)$, and x_k is the mole fraction of residue type k.

RAPDF:

$$N_{\text{exp}}(i,j,r) = \frac{N_{\text{obs}}(i,j)}{N_{\text{total}}} N_{\text{obs}}(r) \tag{3}$$

where $N_{\text{obs}}(i,j) \equiv \sum_r N_{\text{obs}}(i,j,r)$ $N_{\text{obs}}(r) \equiv \sum_{i,j} N_{\text{obs}}(i,j,r)$ and $N_{\text{total}} \equiv \sum_{i,j,r} N_{\text{obs}}(i,j,r)$

DFIRE:

$$N_{\text{exp}}(i,j,r) = (r/r_{\text{cut}})^{\alpha}(\Delta r/\Delta r_{\text{cut}})N_{\text{obs}}(i,j,r_{\text{cut}}) \tag{4}$$

where α is adopted 1.57. The data set for the statistics came from culledpdb (http://dunbrack.fccc.edu/PISCES.php) and the Protein Data Bank (PDB), a total of 2067 entries were included with the percentage identity <30% and X-ray crystal structure resolution better than 2.0 Å. In this work, the first Δr is 0–3 Å, the distance 3–15 Å is binned every 1 Å, resulting total of 25 intervals.

3 RESULTS AND DISCUSSION

3.1 Prediction performance

We applied the method described above to the AmylHex data set. The ability of the hexapeptides from the AmylHex data set to form fibrils is known and can therefore be used to examine our prediction performance. Figure 2 shows *receiver operating characteristics* (ROC) curves based on the residue-based statistical potentials of KBP, DFIRE and RAPDF. Here, 'sensitivity' is defined as the positive fraction of prediction for fibril-forming peptides, and 'specificity' as the negative fraction of prediction for non-fibril-forming peptides. KBP and DFIRE gave almost identical results and were better suited for further study than RAPDF. The different behavior of the three statistical potentials may come from their different reference state definition. We used the statistical potential KBP for further energy score evaluations in the following study.

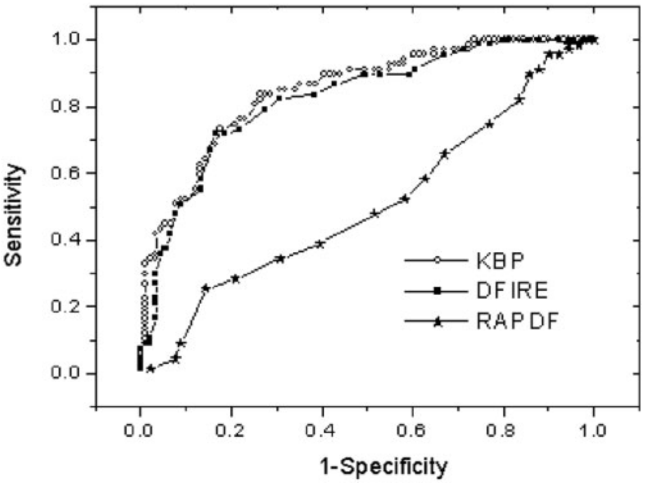

Fig. 2. ROC curves based on statistical potentials of KBP, DFIRE and RAPDF. ROC curves were obtained using the AmylHex sequence sets.

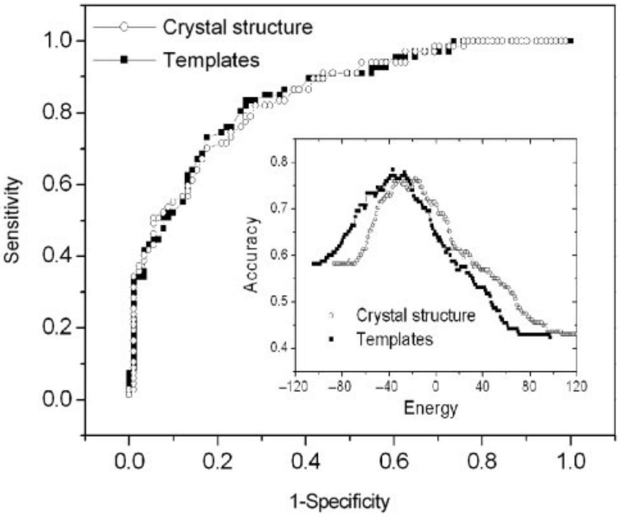

Fig. 3. ROC curve based on the statistical potential of KBP by using the AmylHex sequence sets. The insert plots accuracy as a function of energy score. The open circle curve was obtained from the native microcrystal structure of the NNQQNY peptide from yeast Sup35 and the solid square curve from perturbed template structures.

Figure 3 (○) shows the ROC curve when using the single template of the native microcrystal structure of NNQQNY. In comparison with the perturbed template library (■), we found that there was no significant improvement when using the template structures. The predicted accuracy (defined as the fraction of correct prediction for both fibril-forming and non-fibril-forming sets) in the inserted graph indicates that the curve for the application of the template structures shifts toward lower energy due to the adjustment of the arrangement of the side chains to give a more rational packing structure. The two accuracy curves are very similar in shape and maxima (78% for templates and 77% for microcrystal structure), implying that the prediction performance is not very sensitive to the refinement of the stacking structure when using the residue-based statistical potential. Therefore, these data suggest that both algorithms can be used to predict the amyloid-forming propensity of peptides, and that the single template method may reserve more computational time for larger systems.

The difference between the predictions using our method and that described by Thompson *et al.*, in which the physical-based potential ROSETTADESIGN was used, was relatively small. For example, when 1-specificity is 60%, the sensitivity is 96% in our study, and 100% in their study; when 1-specificity is 14%, the sensitivity is 64% in our study, and 69% in their study. Besides, when using the energy threshold –27 corresponding to the maximum accuracy (which also corresponds to the minimum *P*-value calculated as described in Thompson *et al.* work), where 83.6% of the positives were collected and 26.4% of the negatives were accumulated, the false-positive error rate [defined as 'false positive/(false positive + true positive)'] was 30%, which is between 45% obtained in the 'permissive' threshold and 22% obtained in the 'conservative' threshold in their study. Therefore, we believe that the residue-based statistical potential used here is capable of prediction performance comparable to their method using physical-based atomic level potential ROSETTADESIGN. Furthermore, our method will save much computational power because the residue-based statistical potential simplifies the calculation by using one point C_β to represent one amino acid residue.

We next used another hexapeptide set AmylFrag and control set to test if our method could identify fibril-forming segments of proteins. AmylFrag set was obtained from those fragments confirmed to form fibrils in experiments, while the peptides in control set were unknown for their fibril-forming ability. Figure 4 shows that the AmylFrag curve shifts to lower energy compared to the control curve distinctly. A ratio of their value is approximate 2 at the energy threshold of −27, and reached a maximum of more than 4 at the energy threshold of −66. In Thompson *et al.*'s work, this ratio peaks around the conservative threshold with a value about 2.25–2.5. Therefore, the results shown in both Figures 3 and 4 demonstrate a comparable ability to predict fibril-formation using our method as compared to the method described by Thompson *et al.*

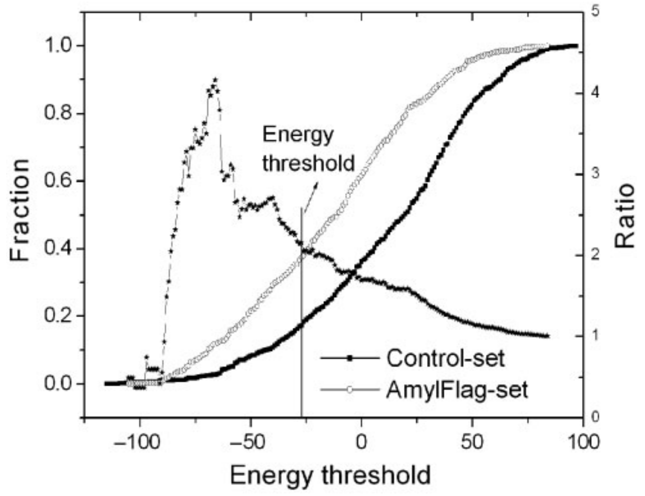

Fig. 4. Fraction of the sequences with energy below or equal to the value at the *x* axis for AmylFrag set and control set. The line + star curve (read from the *y* axis on the right) denotes the ratio of these two fractions, and the vertical line represents the position at energy threshold −27.

3.2 Identification of fibril formation regions of several proteins

We first tested the sequence set studied by Ivanova *et al.* (Ivanova *et al.*, 2006), which included 59 six-residue peptides (16 from the insulin sequence and 43 from the β_2-microglobulin sequence); eight of these sequences formed amyloid-like fibrils or needles in their experiments. Our results show that 10 of the 59 hexamers have strong propensity to form fibrils (where their energy was lower than the energy threshold −27), 4 of which were identified in Ivanova *et al.*'s experiments, as shown in Table 1. Table 1 also lists the hexamers predicted by the method of Thompson *et al.*'s. There are five hexapeptides predicted in our study identical to those identified by their study. Although our approach predicted less than theirs in the fibril-forming peptides confirmed by experiments, the false prediction number is also smaller. The detailed analysis of β_2-microglobulin is described in a subsequent section.

We also applied our method to predict several β-rich proteins to test if it can distinguish fibril-forming fragments from those beta-strand sheet-forming only fragments (Supplementary Fig. S1). Many of the β strands forming peptides are not within the predicted fibril-forming regions, and some of the predicted fibril-forming regions are not in β structure. As there is no report about whether these proteins could form fibrils, there might be false positive predictions. For proteins known to for fibril, this method could identify fibril-forming regions. However, running on unknown proteins may give false positives.

Table 1. Prediction results of the 59 hexapeptides used in this work and in Thompson *et al.*'s study (Thompson *et al.*, 2006), as well as those observed to form amyloid-like fibrils or needles in experiments [(Ivanova *et al.*, 2006)

Hexamer sequence predicted in this study	Hexamer sequence predicted in the reference work	Amyloid-like fibrils or needles observed in experiments
WSFYLL	SLYQLE	
EALYLV	YQLENY	
GFFYTP	SHLVEA	
RGFFYT	KVEHSD	
FFYTPK	LSFSKD	
YVSGFH	**YVSGFH**	
FYLLYY	**FYLLYY**	√
VEALYL	**VEALYL**	√
LLYYTE	**LLYYTE**	√
NHVTLS	**NHVTLS**	√
	KDWSFY	√
	LVEALY	√
	KIVKWD	√
	LYQLEN	√
	FVNQHL	
	GSHLVE	
	HLVEAL	
	SDLSFS	

The bold denotes the predicted in our study are same with those in reference (Thompson *et al.*, 2006).

We then examined the well-known amyloidosis-related proteins, Aβ (1–42), β2-microglobulin, Islet amyloid protein (IAPP), α-synuclein and transthyretin (TTR). The sequences were scanned with a six-residue window and the predicted results are shown in Figure 5, where the lowest energy obtained from template structures was plotted at the position of the initial residue of each six-residue peptide such that the prediction for each point at x corresponded to residues $x - x + 5$. The predicted regions in this study, the corresponding experimental results, as well as those from other predictions [(a) De Groot *et al.*, 2005; (b) Fernandez-Escamilla *et al.*, 2004; (c) Galzitskaya *et al.*, 2006; (d) Pawar *et al.*, 2005] are together presented in Table 2. Detailed descriptions and discussions for each of these proteins are described below.

3.2.1 Aβ (1–42) The hydrophobic peptide Aβ is the principle component of the extracellular amyloid aggregates found in the brains of Alzheimer's patients (Selkoe, 2001). The most abundant amyloid forms include Aβ40 and Aβ42. Figure 5a shows our fibril-forming prediction profile of the 42mer peptide, in which regions containing residues 15–23 and 29–42 showed a strong propensity to form fibrils. Proline scanning mutagenesis (Williams *et al.*, 2004) indicated that the fragment containing residues 15–21 was particularly important in amyloid formation, and it has also been defined as being part of the core fibril structure by solid-state NMR (Petkova *et al.*, 2002) and site-directed spin labeling (Torok *et al.*, 2002). Similarly, Balbach *et al.* (Balbach *et al.*, 2000) reported that the fragment of residues 16–22 formed fibrils. These fragments were all covered in the first region of our prediction. Experiments suggested that residues 30–40 are located in the core of the fibrils (Petkova *et al.*, 2002), and that residues 34–42 readily form ordered β structures (Lansbury *et al.*, 1995). Moreover, the Aβ42 aggregated into fibrils much more rapidly than Aβ40 (Jarrett *et al.*, 1993), which is also in agreement with our predictions. For the protein Aβ42, the references (a), (c) and (d) gave very similar results with ours, while the reference (b) gave a shorter peptide (32–36) for the second predicted region.

3.2.2 β_2-Microglobulin β_2-Microglobulin is a small (99 residues) non-glycosylated protein related to dialysis amyloidosis (Koch, 1992). Four fragments of this protein were predicted by our method to form fibrils (Fig. 5b, residues 21–32, 60–70, 79–84 and 83–89). Likewise, the ability of segments containing residues 20–41, 59–71 and 83–89 to form fibrils has been experimentally demonstrated (Ivanova *et al.*, 2004; Jones *et al.*, 2003; Kozhukh *et al.*, 2002). The C-terminal residues 72–99 have also been reported to self-assemble into fibrillar structures (Ivanova *et al.*, 2003), such that the fragment containing residues 79–84 may be another core except residues 83–89 for fibril formation in this region. In addition, Hasegawa and co-workers found that residues 21–31 formed amyloid fibrils by themselves and with fragment 78–86, while the N-terminal region (containing residues 6–12) and the C-terminal region (containing residues 91–99) did not, all of which is consistent with our predictions. Therefore, our method accurately predicted the ability of peptide segments in β_2-microglobulin to form fibrils. Besides, the four regions in our result are very similar with those in reference (c). While the prediction in reference (b) gave three of the four regions, and reference (d) gave only two of them.

3.2.3 Islet amyloid protein(IAPP) The IAPP accumulates into fibrils in the pancreas of individuals with type II diabetes (Hoppener *et al.*, 2000). Our method predicted that four fragments of this polypeptide, including residues 2–10, 12–18, 15–20 and 22–28, might be able to form fibrils (Fig. 5c). Experiments confirmed that the short peptides NFLVH (residues 14–18) and FLVHS (residues 15–19) formed amyloid fibrils (Mazor *et al.*, 2002), which are located within the fragments 12–18 and 15–20 that our algorithm predicted. The segment of residues 20–29 also formed protofibrils (Goldsbury *et al.*, 2000) in experiments, which covered the fragment containing residues 22–28 predicted by our method. For fragment 2–10, the two Cys (residue 2 and residue 7) might cause inaccurate predictions since many Cys residues form disulfide bonds which may influence the statistics for residue-based potential. Whether or not this segment forms amyloid fibrils has yet to be investigated. Comparison with the references (a–d) in Table 2, it shows that only one fragment given in reference (b), and two regions predicted in reference (c), all of which were included in our predictions.

3.2.4 α-synuclein α-Synuclein is the major component of fibrils that form in the lewy bodies of Parkinson's disease and other neurodegenerative disorders known as synucleinopothies (Maries *et al.*, 2003). We predicted five segments that might form fibrils including residues 13–20, 35–42, 47–57, 66–79 and 85–96. The segments of residues 3–18 (Bodles and Irvine, 2004) and region 34–101 (Der-Sarkissian *et al.*, 2003) have been observed to form β-like aggregates as demonstrated experimentally. Likewise, residues 61–95 (Han *et al.*, 1995) and 69–79 (el-Agnaf and Irvine, 2002) were also suggested or observed to form amyloid fibrils. Together, these data indicate that the fragments we predicted may be the core of the fibrils. In addition, the observation that the fragment containing residues 19–35 (Bodles and Irvine, 2004) and the C-terminus (about 40 amino acids) (Der-Sarkissian *et al.*, 2003) remained unstructured is also consistent with our predictions. For this protein, as shown in Table 2, both in the references (a) and (d), only four of the five fragments predicted in our study were included; the first predicted region (residues 1–8) in reference (c) was not found in any other studies including ours; and no fragment was predicted to be able to form fibrils in reference (b).

3.2.5 Transthyretin (TTR) Transthyretin (TTR), a homotetramer of 127 amino acids, constitutes the fibrillar protein found in familial amyloidotic polyneurophathy (FAP) and senile systemic amyloidosis (SSA) (Saraiva *et al.*, 1984). According to the analysis of the fibril-forming prediction in Figure 5e, the TTR monomer contains several fragments including residues 10–15, 12–17, 25–34, 91–98, 104–113, and 117–124, that may aggregate into fibrils. To date, two regions have been experimentally demonstrated to form amyloid fibrils, one is the segment of residues 10–19 (Jarvis *et al.*, 1994), which includes the residues 10–15 and 12–17 we predicted; the other is region of residues 105–115 (Jaroniec *et al.*, 2002), which also

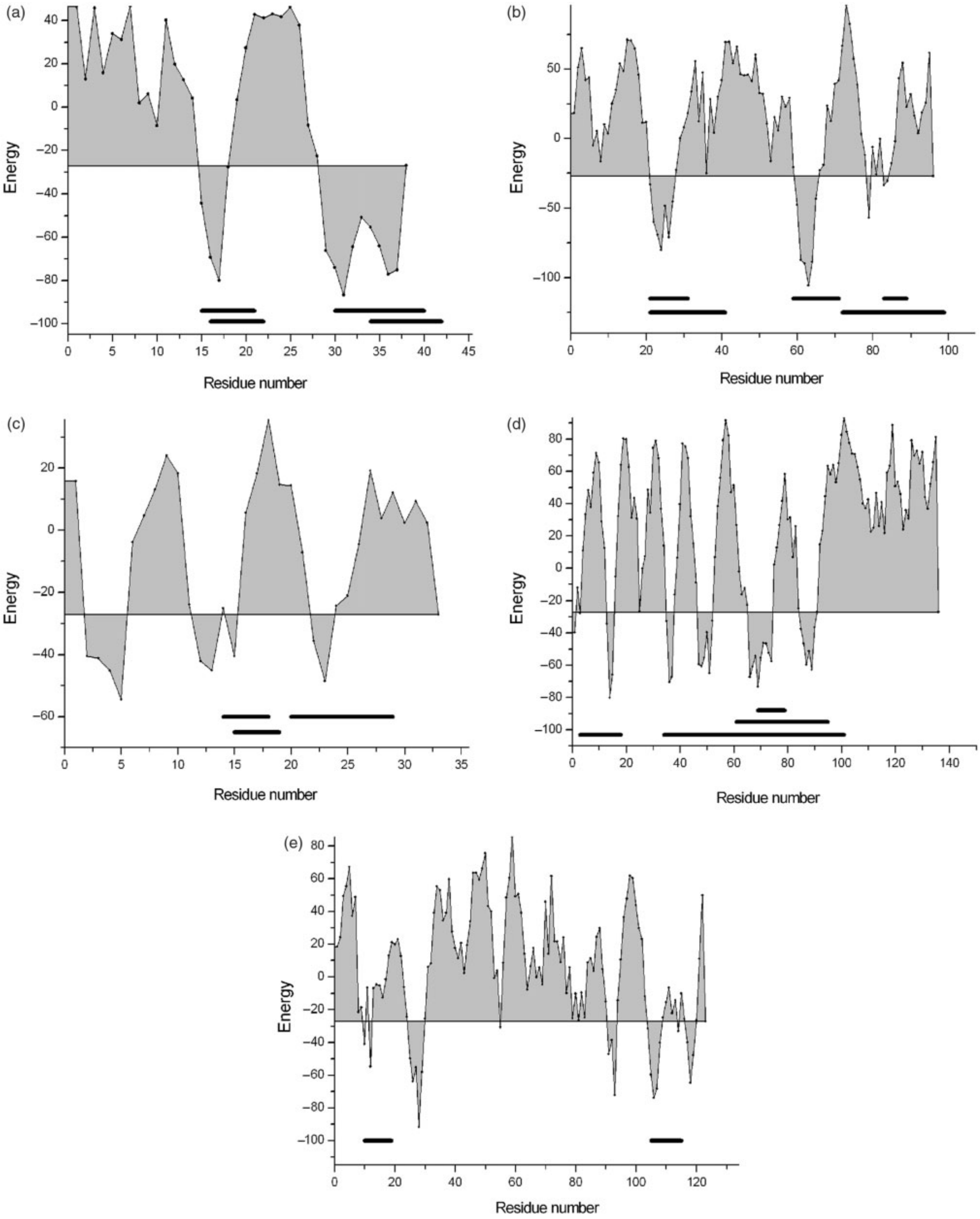

Fig. 5. Fibril-forming prediction profiles of example proteins. (**a**) Aβ (1–42), (**b**) β2-microglobulin, (**c**) Islet amyloid protein (IAPP), (**d**) α-synuclein and (**e**) transthyretin (TTR). The boundary between the fibril-forming and non-forming sections was set at the energy threshold of −27. The black horizontal bars indicate regions which have been found or suggested to form fibrils in experiments.

Table 2. The predicted regions of the five example proteins analyzed in our study and in references (a–d): (a) (Pawar *et al.*, 2005), (b) (Galzitskaya *et al.*, 2006), (c) (De Groot *et al.*, 2005) and (d) (Fernandez-Escamilla *et al.*, 2004), as well as those indicated in experiments

Name of protein (residue number in it)	Amyloidogenic regions experimentally determined	Predicted regions in this study	Predicted regions in Ref (a)	Predicted regions in Ref (b)	Predicted regions in Ref (c)	Predicted regions in Ref (d)
Aβ peptide (42)	15–21, 16–22	***15–23***	15–21	16–21	16–21	17–21
	30–40, 34–42	***29–42***	30–42	32–36	30–42	29–42
β_2-Microglobulin (100)	21–31, 21–41	***21–32***	ND	22–29	21–31	18–30
	59–71	***60–70***		60–69	56–69	60–72
	72–99, 83–89	***79–84,83–89***		82–86	79–85, 87–91	
Islet amyloid protein, IAPP (37)	14–18, 15–19	2–10	ND			
	20–29	***12–18, 15–20***		13–18	12–18	
		22–28			22–28	
α-Synuclein (140)	3–18, 34–101, 61–95, 69–79	***13–20***			1–8	
		35–42	38–40		27–56	15–19, 36–41
		47–57	50–54			52–55
		66–79	65–75		61–94	69–78
		85–96	87–92			
Transthyretin, TTR (127)	10–19	***10–15, 12–17***	ND	11–16	10–20	
		25–34, 91–98		27–34, 77–81	23–33	
	105–115	***104–113***		105–110	105–118	106–111
		117–124				118–122

The bold and italic characters denote the regions predicted in this study also were observed experimentally. 'ND' represents that there are no data for this protein in the reference.

was identified in this study. Table 2 displays that the region 10–19 was also predicted in references (b) and (c). Meanwhile, the region 105–115 was given in all other reference [except reference (a) as there were no data for this protein].

3.3 Prediction profiles of the yeast proteome

We tried to analysis the entire yeast proteome using the protein sequence set from UniProt/TrEMBL in EBI. The set contains the amino acid sequences of 5869 translated from systematically identified ORFs (open reading frames), and excludes dubious ORFs and pseudogenes. By scanning the sequences with a six-residue window, and excluding those peptides contained cysteine residues, a total of 2 461 399 unique hexapeptides were obtained. Analysis of this hexamer set predicted that 361 778 hexapeptides would form amyloid fibrils, which accounts for ~1.59% of the entire set. The highest 180 segments are listed in the Supplementary Material (Table S2), from which it is obvious that most of them contain hydrophobic residues such as isoleucine, phenylalanine, valine, leucine and tryptophan. These data indicate that a hydrophobic core is very favorable for fibril formation.

4 CONCLUSION

In this study, we used the microcrystal structure of the hexapeptide NNQQNY to demonstrate that the residue-based statistical potential can be used to identify amyloidogenic fragments of peptides and proteins. Our predictions were comparable to those based on the physical-based potential ROSETTADESIGN. Examination of the proteins and peptides related to well-known amyloidosis agreed with experimental data. As the residue-based statistical potential calculations are computationally efficient, this structure-based approach can be used to analyze large systems, such as entire proteomes.

The major limitation of the current method lies in that only the microcrystal X-ray structure of NNQQNY was used, which does not show common fibril twist and can not represent structure types other than parallel stacking within β-sheet. We expect that more experimental structure models for amyloid fibril will improve the predictive power of the current method.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Prof. Yaoqi Zhou for helpful suggestions and discussions. This work was supported in part by the National Key Basic Research Project of China (No. 2003CB715900) and National Natural Science Foundation of China (90403001, 30490245). The authors thank Peking University Center for Computational Science & Engineering for the use of its computational facility.

Conflict of Interest: none declared.

REFERENCES

Balbach,J.J. *et al.* (2000) Amyloid fibril formation by A beta 16-22, a seven-residue fragment of the Alzheimer's beta-amyloid peptide, and structural characterization by solid state NMR. *Biochemistry*, **39**, 13748–13759.

Balbirnie,M. *et al.* (2001) An amyloid-forming peptide from the yeast prion Sup35 reveals a dehydrated beta-sheet structure for amyloid. *Proc. Natl Acad. Sci. USA*, **98**, 2375–2380.

Bemporad,F. *et al.* (2006) Sequence and structural determinants of amyloid fibril formation. *Acc. Chem. Res.*, **39**, 620–627.

Bodles,A.M. and Irvine,G.B. (2004) Alpha-synuclein aggregation. *Protein Pept. Lett.*, **11**, 271–279.

Bowie,J.U. *et al.* (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.

Caflisch,A. (2006) Computational models for the prediction of polypeptide aggregation propensity. *Curr. Opin. Chem. Biol.*, **10**, 437–444.

De Groot,N. *et al.* (2005) Prediction of 'hot spots' of aggregation in disease-linked polypeptides. *BMC Struct. Biol.*, **5**, 18.

Der-Sarkissian,A. *et al.* (2003) Structural organization of alpha-synuclein fibrils studied by site-directed spin labeling. *J. Biol. Chem.*, **278**, 37530–37535.

Dobson,C.M. (1999) Protein misfolding, evolution and disease. *Trends Biochem. Sci.*, **24**, 329–332.

Dobson,C.M. (2001) The structural basis of protein folding and its links with human disease. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, **356**, 133–145.

Dobson,C.M. (2002) Getting out of shape. *Nature*, **418**, 729–730.

el-Agnaf,O.M. and Irvine,G.B. (2002) Aggregation and neurotoxicity of alpha-synuclein and related peptides. *Biochem. Soc. Trans.*, **30**, 559–565.

Fandrich,M. *et al.* (2001) Amyloid fibrils from muscle myoglobin. *Nature*, **410**, 165–166.

Fernandez-Escamilla,A.M. *et al.* (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.*, **22**, 1302–1306.

Galzitskaya,O.V. *et al.* (2006) Is it possible to predict amyloidogenic regions from sequence alone? *J. Bioinform. Comput. Biol.*, **4**, 373–388.

Goldsbury,C. *et al.* (2000) Amyloid fibril formation from full-length and fragments of amylin. *J. Struct. Biol.*, **130**, 352–362.

Guijarro,J.I. *et al.* (1998) Amyloid fibril formation by an SH3 domain. *Proc. Natl Acad. Sci. USA*, **95**, 4224–4228.

Han,H. *et al.* (1995) The core Alzheimer's peptide NAC forms amyloid fibrils which seed and are seeded by beta-amyloid: is NAC a common trigger or target in neurodegenerative disease? *Chem. Biol.*, **2**, 163–169.

Hoppener,J.W. *et al.* (2000) Islet amyloid and type 2 diabetes mellitus. *N. Engl. J. Med.*, **343**, 411–419.

Ivanova,M.I. *et al.* (2003) Role of the C-terminal 28 residues of beta2-microglobulin in amyloid fibril formation. *Biochemistry*, **42**, 13536–13540.

Ivanova,M.I. *et al.* (2004) An amyloid-forming segment of beta2-microglobulin suggests a molecular model for the fibril. *Proc. Natl Acad. Sci. USA*, **101**, 10584–10589.

Ivanova,M.I. *et al.* (2006) A systematic screen of beta(2)-microglobulin and insulin for amyloid-like segments. *Proc. Natl Acad. Sci. USA*, **103**, 4079–4082.

Jaroniec,C.P. *et al.* (2002) Molecular conformation of a peptide fragment of transthyretin in an amyloid fibril. *Proc. Natl Acad. Sci. USA*, **99**, 16748–16753.

Jarrett,J.T. *et al.* (1993) The carboxy terminus of the beta amyloid protein is critical for the seeding of amyloid formation: implications for the pathogenesis of Alzheimer's disease. *Biochemistry*, **32**, 4693–4697.

Jarvis,J.A. *et al.* (1994) 1H NMR analysis of fibril-forming peptide fragments of transthyretin. *Int. J. Pept. Protein Res.*, **44**, 388–398.

Jones,D.T. *et al.* (1992) A new approach to protein fold recognition. *Nature*, **358**, 86–89.

Jones,S. *et al.* (2003) Amyloid-forming peptides from beta2-microglobulin-Insights into the mechanism of fibril formation in vitro. *J. Mol. Biol.*, **325**, 249–257.

Klunk,W.E. *et al.* (1999) Quantifying amyloid by congo red spectral shift assay. *Meth. Enzymol.*, **309**, 285–305.

Koch,K.M. (1992) Dialysis-related amyloidosis. *Kidney Int.*, **41**, 1416–1429.

Kozhukh,G.V. *et al.* (2002) Investigation of a peptide responsible for amyloid fibril formation of beta 2-microglobulin by achromobacter protease I. *J. Biol. Chem.*, **277**, 1310–1315.

Kuhlman,B. and Baker,D. (2000) Native protein sequences are close to optimal for their structures. *Proc. Natl Acad. Sci. USA*, **97**, 10383–10388.

Lansbury,P.T. *et al.* (1995) Structural model for the beta-amyloid fibril based on interstrand alignment of an antiparallel-sheet comprising a C-terminal peptide. *Nat. Struct. Biol.*, **2**, 990–998.

Lopez De La Paz,M. *et al.* (2002) De novo designed peptide-based amyloid fibrils. *Proc. Natl Acad. Sci. USA*, **99**, 16052–16057.

Lopez de la Paz,M. and Serrano,L. (2004) Sequence determinants of amyloid fibril formation. *Proc. Natl Acad. Sci. USA*, **101**, 87–92.

Lu,H. and Skolnick,J. (2001) A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins*, **44**, 223–232.

Maries,E. *et al.* (2003) The role of alpha-synuclein in Parkinson's disease: insights from animal models. *Nat. Rev. Neurosci.*, **4**, 727–738.

Mazor,Y. *et al.* (2002) Identification and characterization of a novel molecular-recognition and self-assembly domain within the islet amyloid polypeptide. *J. Mol. Biol.*, **322**, 1013–1024.

Nelson,R. *et al.* (2005) Structure of the cross-beta spine of amyloid-like fibrils. *Nature*, **435**, 773–778.

Pawar,A.P. *et al.* (2005) Prediction of "aggregation-prone" and "aggregation-susceptible" regions in proteins associated with neurodegenerative diseases. *J. Mol. Biol.*, **350**, 379–392.

Petkova,A.T. *et al.* (2002) A structural model for Alzheimer's beta -amyloid fibrils based on experimental constraints from solid state NMR. *Proc. Natl Acad. Sci. USA*, **99**, 16742–16747.

Rochet,J.C., Lansbury,P.T. and Jr. (2000) Amyloid fibrillogenesis: themes and variations. *Curr. Opin. Struct. Biol.*, **10**, 60–68.

Ross,C.A. and Poirier,M.A. (2004) Protein aggregation and neurodegenerative disease. *Nat. Med.*, **10** (Suppl.), S10–S17.

Rousseau,F. *et al.* (2006) How evolutionary pressure against protein aggregation shaped chaperone specificity. *J. Mol. Biol.*, **355**, 1037–1047.

Saiki,M. *et al.* (2006) Interaction-based evaluation of the propensity for amyloid formation with cross-beta structure. *Biochem. Biophys. Res. Commun.*, **343**, 1262–1271.

Samudrala,R. and Moult,J. (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.*, **275**, 895–916.

Saraiva,M.J. *et al.* (1984) Amyloid fibril protein in familial amyloidotic polyneuropathy, Portuguese type. Definition of molecular abnormality in transthyretin (prealbumin). *J. Clin. Invest.*, **74**, 104–119.

Selkoe,D.J. (2001) Alzheimer's disease: genes, proteins, and therapy. *Physiol. Rev.*, **81**, 741–766.

Skolnick,J. *et al.* (1997) MONSSTER: a method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol.*, **265**, 217–241.

Tanaka,S. and Scheraga,H.A. (1976) Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*, **9**, 945–950.

Tenidis,K. *et al.* (2000) Identification of a penta- and hexapeptide of islet amyloid polypeptide (IAPP) with amyloidogenic and cytotoxic properties. *J. Mol. Biol.*, **295**, 1055–1071.

Thompson,M.J. *et al.* (2006) The 3D profile method for identifying fibril-forming segments of proteins. *Proc. Natl Acad. Sci. USA*, **103**, 4074–4078.

Torok,M. *et al.* (2002) Structural and dynamic features of Alzheimer's Abeta peptide in amyloid fibrils studied by site-directed spin labeling. *J. Biol. Chem.*, **277**, 40810–40815.

Ventura,S. *et al.* (2004) Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case. *Proc. Natl Acad. Sci. USA*, **101**, 7258–7263.

Williams,A.D. *et al.* (2004) Mapping abeta amyloid fibril secondary structure using scanning proline mutagenesis. *J. Mol. Biol.*, **335**, 833–842.

Yoon,S. and Welsh,W.J. (2004) Detecting hidden sequence propensity for amyloid fibril formation. *Protein Sci.*, **13**, 2149–2160.

Zhou,H. and Zhou,Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.*, **11**, 2714–2726.