Expected packing density allows prediction of both amyloidogenic and disordered regions in protein chains

2007 J. Phys.: Condens. Matter 19 285225

(http://iopscience.iop.org/0953-8984/19/28/285225)

IOP PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **19** (2007) 285225 (15pp) doi:10.1088/0953-8984/19/28/285225

Expected packing density allows prediction of both amyloidogenic and disordered regions in protein chains

Oxana V Galzitskaya[1], **Sergiy O Garbuzynskiy and Michail Yu Lobanov**

Institute of Protein Research, Russian Academy of Sciences, 142290, Pushchino, Moscow Region, Russia

E-mail: ogalzit@vega.protres.ru

Received 29 September 2006, in final form 17 October 2006
Published 25 June 2007
Online at stacks.iop.org/JPhysCM/19/285225

Abstract

The determination of factors that influence conformational changes in proteins is very important for the identification of potentially amyloidogenic and disordered regions in polypeptide chains. In our work we introduce a new parameter, mean packing density, to detect both amyloidogenic and disordered regions in a protein sequence. It has been shown that regions with strong expected packing density are responsible for amyloid formation. Our predictions are consistent with known disease-related amyloidogenic regions for 9 of 12 amyloid-forming proteins and peptides in which the positions of amyloidogenic regions have been revealed experimentally. Our findings support the concept that the mechanism of formation of amyloid fibrils is similar for different peptides and proteins. Moreover, we have demonstrated that regions with weak expected packing density are responsible for the appearance of disordered regions. Our method has been tested on datasets of globular proteins and long disordered protein segments, and it shows improved performance over other widely used methods. Thus, we demonstrate that the expected packing density is a useful value for predicting both disordered and amyloidogenic regions of a protein based on sequence alone. Our results are important for understanding the structural characteristics of protein folding and misfolding.

1. Introduction

The formation of amyloid fibrils is associated with an increase in β-structure content, which leads to fibrillar aggregation [1]. In addition to proteins observed in amyloid diseases, recent studies have shown that diverse proteins not related to any amyloid disease can aggregate into fibrils under destabilizing conditions [2–4]. Normal proteins can become toxic when they

[1] Author to whom any correspondence should be addressed.

undergo fibrillation [5]. Therefore, the mechanism of amyloid formation is under intensive investigation. Recognition of the factors that influence conformational changes and misfolding in proteins is a general fundamental problem, the solution of which will help in the search for effective treatments for amyloid illnesses.

The experimental observation that specific continuous regions of amyloid-forming proteins are more amyloidogenic than others suggests that there is a sequence propensity for amyloid formation. Moreover, the observation that some short peptides can also form amyloids implies that exposure of short segments of proteins can nucleate the transition of native proteins into the amyloid state and suggests that fibril formation is sequence specific [6]. In the mechanism of amyloidogenesis for natively folded proteins such as β_2-microglobulin and transthyretin, the observed partial unfolding is believed to be a prerequisite for the assembly of the proteins into amyloid fibrils both *in vitro* and *in vivo* [7]. It has been suggested that residues with enhanced flexibility and accessibility to solvent are important for the initiation of fibrillation [8]. This means that partial unfolding of the rigid native structure can provide a specific interface for the beginning of fibrillation. Thus, to understand the molecular mechanism of amyloidosis, it is necessary to find factors that induce partial unfolding of proteins and subsequent formation of amyloid fibrils at or near physiological conditions.

Some natively unfolded proteins are involved in amyloid diseases (type II diabetes, Alzheimer's and Parkinson's diseases). Nevertheless, most of the natively unfolded proteins do not undergo aggregation [9]. This fact indicates that unfolding is a necessary but not sufficient condition for aggregation. Knowledge of characteristics that control the process of amyloid fibril formation is important for finding effective drugs for treatment of amyloid diseases.

The first high-resolution (1 Å) crystal of an amyloid fibre formed by a sequence-designed polypeptide has been obtained [10]. Recently, the atomic structure of the cross-β spine [11] for a seven-residue peptide segment from Sup35 (GNNQQNY) was determined. It is a double β-sheet, in which each sheet is formed from parallel segments stacked in register. Side chains protruding from the two sheets form a dry, tightly self-complementing steric zipper that bonds the sheets. Within each sheet, every segment is bound to two neighbouring segments through stacks of both backbone and side-chain hydrogen bonds.

There are several computational methods for predicting a protein's propensity to amyloid fibril formation. In the work of Fernandez *et al* [12] it has been shown that a concentration of such defects as insufficient shielding of hydrogen bonds from attack by water might yield an aggregation-induced nucleus. But the analysis of these defects revealed that extensive exposure of hydrogen bonds to attack by water might be a necessary but not sufficient condition to imply a propensity for organized aggregation [12].

A computational algorithm has been suggested that detects the non-native (hidden) β-strand propensity of sequences by considering the relationships between protein local sequence and secondary structure in terms of tertiary contacts [13]. This algorithm detects sequences within the protein that are favourable for triggering the formation of amyloid fibrils. It is worthwhile emphasizing here that both algorithms for prediction of the amyloidogenic properties of polypeptide chains that are considered above can be applied only to those proteins for which the three-dimensional structure is known.

On the other hand, there is a method for the prediction of amyloidogenic regions from the amino acid sequence alone [14]. After the experimental investigation of the amyloidogenic properties of a model six-residue peptide and its mutants, the authors obtained a six-residue amyloidogenic pattern (STVIIE) and used this pattern for the identification of amyloidogenic fragments in proteins [14]. This amyloidogenic pattern has been used to validate the premise that the amyloidogenicity of a protein is indeed localized in short protein stretches (the amyloid

stretch hypothesis [15]). It has been demonstrated that the conversion of a soluble non-amyloidogenic protein (SH3 domain of α-spectrin) into an amyloidogenic-prone molecule can be triggered by a non-destabilizing six-residue amyloidogenic insertion in a particular structural environment.

Recently, a new method for identifying the fibril-forming segments of proteins has been suggested [16]. This method is based on the threading of six-residue peptides through the known crystal structure of an amyloid fibre [11] formed by the peptide from Sup35. The putative prediction is accepted as a prediction if its energy evaluated with ROSETTADESIGN (www.rosettacommons.org) is lower than the threshold energy.

The formation of a sufficient number of interactions is necessary to compensate for the loss of conformational entropy during the protein folding process. Therefore, the structural uniqueness of native proteins is a result of the balance between the conformational entropy and the energy of residue interactions. It seems that disordered regions in a protein chain do not have a sufficient number of interactions to compensate for the loss of conformational entropy that results from the formation of a globular state. On the other hand, a large increase in the energy of interactions will lead to a loss of the unique structure because strengthening of contact energy will speed up folding, but it is also likely to lead to erroneous folds (for example, to amyloid fibrils).

It has been suggested that the lack of a rigid globular structure under physiological conditions might represent a considerable functional advantage for 'natively unfolded' proteins. Their large plasticity allows them to interact efficiently with several different targets compared to a folded protein with limited conformational flexibility [17–20]. It has been shown that disordered regions are involved in DNA binding and other types of molecular recognition [21]. A large portion of the sequences of 'natively unfolded' proteins contain segments of low complexity and high predicted flexibility [22–29]. It also has been indicated that a combination of low overall hydrophobicity and a large net charge represents a structural feature of 'natively unfolded' proteins in comparison with small globular proteins [9, 30]. There are currently several widely used methods for the prediction of disordered regions: GlobPlot [31] is a simple propensity-based approach evaluating the tendency of residues to be in a regular secondary structure; PONDR VL3H [28] was trained to distinguish experimentally verified disordered proteins from globular proteins by various machine learning approaches; in developing DISOPRED [32] the definition of disorder was restrained to regions that are missing from x-ray structures, and a support vector machine was trained to specifically recognize these; IUPred [33] assigns the order/disorder status to residues on the basis of their ability to form favourable pairwise contacts. We were the first to use the number of contacts per residue as a parameter to distinguish folded and natively unfolded proteins [34]. We have extended our method to predict disordered regions and have made comparisons with the above mentioned methods [35]. It has been demonstrated that our method is the best of the widely used methods.

Despite considerable efforts to understand it, the nature of the appearance of amyloidogenic and unfolded regions remains unclear. The goal of this work is to test our hypothesis about whether protein regions that possess expected strong packing density can be responsible for the amyloidogenic properties of proteins, while regions with weak packing density simultaneously are responsible for the appearance of unfolded regions. We introduce a new parameter, namely mean packing density (number of residues within a given distance from the considered residue), which enables the prediction of both amyloidogenic and unfolded regions from the protein sequence. These findings support the concept that the nature of the appearance of amyloidogenic and unfolded regions has a similar basis in different peptides and proteins.

Table 1. Mean observed packing density for 20 amino acid residues (and errors in determination of average) obtained using contact radius 8.0 Å.

Amino acid residue	Gly	Asp	Pro	Glu	Lys
Number of close residues	17.11 ± 0.02	17.41 ± 0.03	17.43 ± 0.03	17.46 ± 0.02	17.67 ± 0.02
Amino acid residue	Ser	Asn	Gln	Thr	Ala
Number of close residues	18.19 ± 0.03	18.49 ± 0.03	19.23 ± 0.04	19.81 ± 0.03	19.89 ± 0.02
Amino acid residue	Arg	His	Cys	Val	Met
Number of close residues	21.03 ± 0.03	21.72 ± 0.05	23.52 ± 0.05	23.93 ± 0.03	24.82 ± 0.06
Amino acid residue	Leu	Ile	Tyr	Phe	Trp
Number of close residues	25.36 ± 0.02	25.71 ± 0.03	25.93 ± 0.04	27.18 ± 0.04	28.48 ± 0.07

2. Materials and methods

2.1. Observed packing density for 20 types of amino acid residues

The set of protein structures used for calculation of the packing density observed in protein structures was obtained by inspection of the SCOP (structural classification of proteins) database release 1.61 [36]. In all, 5829 domains from four general classes (a–d) with less than 80% sequence identity values were found: 1133 all-α proteins from class a, 1644 all-β proteins from class b, 1617 α/β proteins from class c and 1435 $\alpha + \beta$ proteins from class d. The observed packing density for each amino acid residue from this database was calculated as the number of close residues (within the given distance). In our case a residue is considered close to the given residue if any pair of their heavy atoms is at distance of less than 8 Å. The neighbouring residues bound with peptide bonds (which are close in any case) are not taken into account. The mean observed packing density for each of 20 types of amino acid residues is presented in table 1. These 20 values were used for the prediction of packing density from protein sequences, that is, the expected packing density (we consider the expected packing density of a residue to be equal to the mean observed packing density of the corresponding residue in a globular state).

2.2. Calculation of the expected packing density profile

It is worthwhile emphasizing that the order of the residues may play an important role in protein folding and may account for regions with weak and strong packing density in a protein structure. To predict such regions in a protein, we construct a profile of the expected packing density for the protein sequence. The calculations are based on a sliding window averaging technique. For each peptide and protein, in prediction of amyloidogenic regions the sliding window size is five residues (the smallest experimentally obtained amyloidogenic fragment in disease-related amyloidogenic proteins in our database, see table 2 below) while the sliding window size is 11 (or 41) residues in the case of prediction of unfolded regions. The packing density profile is calculated as follows. First, the expected packing density is determined for each residue (see table 1); then, these numbers are averaged for five residues inside the window and assigned to the central residue of the window. Therefore, the influence of residues along the sequence flanking each window is included in our calculation. The value of the average expected packing density for every position of the polypeptide chain provides the packing density profile. If more than five residues in a row have values over a specified threshold, this region is predicted to be amyloidogenic. On the other hand, any region having more than 11 (or 41) residues with values below a specified threshold is predicted as natively unfolded.

Table 2. Predicted versus experimentally observed amyloid-forming regions in amyloidogenic proteins and peptides.

Name of protein and number of residues in it	Associated pathology	Experimentally investigated amyloidogenic regions	Predicted regions[a]			
			With strong expected packing density	With high hydrophobicity	With high β-propensity	Hybrid scale
τ protein 441	Alzheimer's disease	306–311 [42]	**307–311**	113–117 246–251	77–82 126–130	—
Human prion 253	Creutzfeldt–Jakob disease, fatal familial insomnia, Gerstmann–Straussler disease, Huntington disease-like 1, kuru	132–160 [46] 178–193 [45]	5–15 **136–141** **147–152** 174–178 **180–184** 211–217 240–253	4–15 120–124 127–132 211–216 231–253	8–15 **181–193** 214–218 242–250	5–15 211–216 241–253
Apolipoprotein A-I 243	Involved in hereditary systemic amyloidosis	1–93 [47]	**16–21** **69–74** 113–117 227–231	**15–21** 218–223 227–232	**16–20** **52–59** 199–204 223–231	**16–20** 227–231
Lysozyme 130	Autosomal dominant hereditary amyloidosis	49–64 [48]	25–33 **55–59** **61–65** 107–114	26–32 75–84 126–130	—	26–32
Transthyretin 127	Senile systemic amyloidosis and familial amyloid polyneuropathy	10–19 [50] 105–115 [52]	**11–16** 27–34 77–81 **105–110**	**10–16** 26–32 92–96 **107–113**	30–34 116–122	**106–110**
β_2-microglobulin 99	Dialysis related amyloidosis	20–41 [53] 59–71 [54] 83–89 [6]	**22–29** **60–69** **82–86**	**23–29** **61–68**	**62–69**	**23–28** **61–69**
Amyloid A protein (AA) 76	Reactive amyloidosis	1–11 [55]	**1–6** 16–20 66–70	—	**1–5**	**2–6**
Medin 50	Aortic medial amyloid	42–49 [56]	9–13 20–24	9–14 **44–50**	—	—
$A\beta$ peptide 42	Alzheimer's disease	14–23 [57] 30–38 [58]	**16–21** **32–36**	**32–42**	—	—
Amylin (islet amyloid protein, hIAPP) 37	Type II diabetes	14–19 [60] 20–27 [59]	**13–18**	**14–18**	4–8	**14–18**
NAC peptide of α-synuclein 35	Parkinson's disease Alzheimer's disease	3–18 [62]	—	**8–13**	—	—
Calcitonin 32	Medullary carcinoma thyroid	15–19 [63]	6–11	5–11 27–31	—	6–11

[a] The predicted amyloidogenic regions which have intersections with the experimentally observed ones are indicated in bold.

2.3. Databases used to test our method

To evaluate the accuracy of, and confidence in, our method of predicting amyloidogenic regions, a database of six-residue peptides (67 peptides that form fibrils and 91 peptides that do not form fibrils) was used [16]. To test our method, we also used the amino acid sequences of 12 disease-related amyloidogenic proteins and peptides (for which the position of amyloidogenic regions is localized experimentally); the sequences were taken from the SWISS-PROT database [37] (http://us.expasy.org/sprot/). To test our method for predicting natively unfolded regions, we used three databases. Two of them were downloaded from the Database of Protein Disorder (DisProt) [38]. The first one consists of sequences of 427 completely intrinsically disordered proteins and disordered fragments. The second database contains 129 natively unfolded proteins. The third database consists of 559 globular proteins without natively unfolded fragments [33]. This database was constructed using Protein Data Bank (PDB) entries from the above work.

2.4. Evaluation of the quality of predictions

To obtain the quality of predictions and to determine thresholds, we calculated true positive and false positive rates and constructed the so-called receiver operator characteristic (ROC) curves. In predictions of unfolded regions, the true positive rate was calculated as the fraction of residues predicted as unfolded over the unfolded set of residues; the false positive rate was the fraction of predicted unfolded residues over the set of folded residues. Similarly, in the case of six-residue peptides that were fibril formers, the true positive rate was calculated as the fraction of peptides predicted as fibril formers in the fibril formers set of peptides while the false positive rate was the fraction of peptides predicted as fibril formers in the set of peptides that are fibril non-formers.

2.5. The other scales for prediction of amyloidogenic regions

Using hydrophobicity and β-sheet propensity scales, we predicted the amyloidogenic regions of the considered proteins and peptides and evaluated the obtained results in a similar way to how we analysed expected packing density. The hydrophobicity scale for 20 types of amino acid residues was taken from the work of Fauchere and Pliska [39]. The β-sheet propensities of the 20 types of amino acid residues in an internal β-sheet position were taken from the work of Minor and Kim [40]. In addition, a hybrid scale was obtained by summation of the three different scales (packing density, hydrophobicity and β-sheet propensity) normalized with equal weights. The original hydrophobicity and β-sheet propensity scales were taken with reversed sign since the most hydrophobic and β-sheet-predisposed amino acid residues have the largest negative values.

3. Results

3.1. Observed mean packing density for 20 types of amino acid residues and expected packing density profiles

First we constructed a database of protein structures and calculated the packing density for each amino acid residue in it. The average packing density observed in protein structures for each of the 20 types of amino acid residues is shown in table 1. These values were considered to be the expected packing density for the amino acid residues of the corresponding type. Further, in a protein or peptide sequence for each amino acid residue the corresponding value from table 1

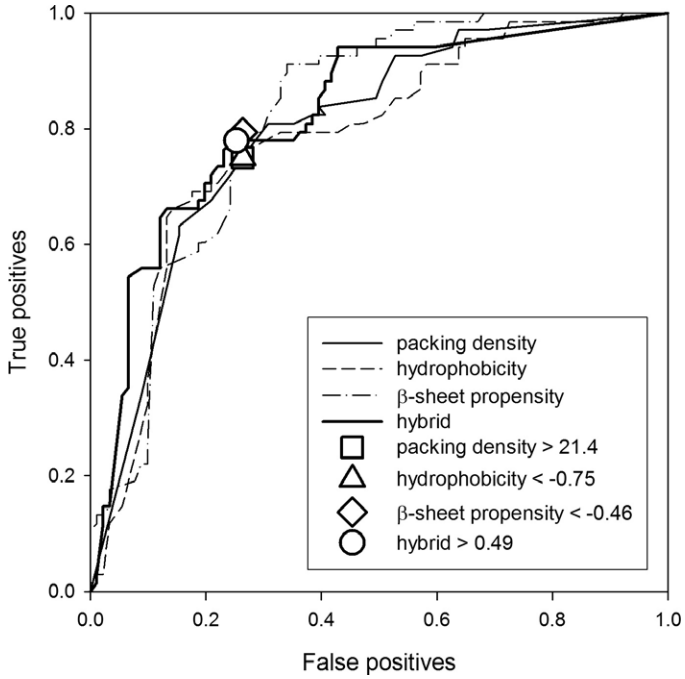

Figure 1. The ROC curves for prediction of amyloidogenic regions in the database of fibril forming and fibril non-forming peptides. The symbols correspond to values chosen as thresholds.

was taken as an expected packing density. The values were averaged over a sliding window, and a packing density profile was produced (see section 2). Similarly, the other types of profiles were built using other scales instead of the scale from table 1 (for example, hydrophobicity profile based on the hydrophobicity scale etc).

3.2. Searching for peptides that are fibril formers and fibril non-formers

To obtain a threshold for our predictions, we took a database of six-residue peptides, some of which were fibril formers and some of which were fibril non-formers [16]. The ROC curves for our method are shown in figure 1. The four ROC curves correspond to four scales: packing density (table 1), hydrophobicity [39], β-sheet propensity [40] and the hybrid scale that was obtained by summation of the three normalized scales (see section 2). For further investigations, we considered the following values as thresholds for predicting amyloidogenic regions (which gave rather a high level of true predictions, about 80%, as well as a rather low level of false predictions, about 25%): packing density greater than 21.4, hydrophobicity less than -0.75, β-sheet propensity less than -0.46 and the value of the hybrid scale greater than 0.49 (the corresponding points on the ROC curves (figure 1) are marked with symbols).

3.3. Searching for amyloidogenic regions in proteins with known disease-related regions

We collected a database of all known proteins and peptides that are associated with amyloid diseases, in which the position of amyloidogenic regions is experimentally examined (see table 2). Amyloids are elongated fibrils that bind the aromatic dyes Congo red and Thioflavin-T and have a common cross-β x-ray diffraction pattern [41].

We constructed a packing density profile for each of these proteins and peptides. As the minimum observed (see table 2) size of amyloidogenic fragments is five residues long, we used a sliding window of five residues (see section 2) and predicted a region as amyloidogenic if five or more sequential residues lie above the considered threshold (the number of close residues within 8 Å is 21.4). Our hypothesis is that regions with a strong expected packing density will probably correspond to aggregation regions, which presumably intersect with amyloidogenic regions of proteins. The experimentally observed amyloidogenic regions and the predicted ones are presented in table 2. One can see that for 9 of 12 examined proteins and peptides the predictions are consistent with the experimentally found amyloidogenic regions.

In Alzheimer's disease, τ-protein forms neurofibrillary tangles, which are bundles of paired helical filaments. A single region (amino acid residues 306–311), which is shown experimentally to be amyloidogenic [42], is correctly predicted by our method (see table 2 and figure 2(a)).

Despite an abundance of experimental data in the search for amyloidogenic regions in human prion protein, it is still difficult to determine which regions these are. It has been shown that helix 1 (residues 144–153) of human prion protein (PrP) plays a critical role in the amyloidogenic process [43, 44]. Peptides corresponding to three helical regions (residues 144–154, helical region one; residues 178–193, helical region two; and residues 198–218, helical region three) have been synthesized and studied [45]. The peptides corresponding to the second helical region, residues 180–193 and residues 178–193, are the only ones that form an amyloid structure, according to data obtained by electron microscopy and Congo red birefringence [45]. By using two intrinsic fluorescent variants of this protein (Y150W and F141W), conformational changes confined to the segment 132–160 were monitored [46]. Our predicted fragments intersect with all helices (see figure 2(b)).

Most mutations described in apolipoprotein A (ApoA) are within the N-terminal portion of the protein (residues 1–93), which represents the proteolysis fragment that is incorporated into amyloid deposits [47]. We predict as amyloidogenic two regions (residues 16–21 and 69–74) within the N-terminal portion as well as two additional regions in the C-terminal part of ApoA, which both have strong expected packing density.

The experimentally found amyloidogenic fragment of lysozyme (residues 49–64), which has been specifically implicated in amyloidogenic conversion [48, 49], is a part of the β-domain in the native structure of the protein. Our predictions for lysozyme are consistent with experimental results (see figure 2(c)); however, two additional fragments (25–33 and 107–114) are also predicted.

The most amyloidogenic peptide fragments from transthyretin (TTR) have been demonstrated in two regions: residues 10–19, which encompass the A strand of the inner β-sheet structure that readily forms amyloid fibrils when dissolved in water at low pH [50, 51], and residues 105–115, which adopt an extended β-strand conformation that is similar to that found in the native protein [52]. We predicted these important regions (11–16 and 105–110) correctly and two additional regions with strong expected packing density (see figure 2(d)).

It has been found experimentally that the following sequences play a dominant role in the amyloidogenesis of β_2-microglobulin: residues 20–41 [53], residues 59–71 [54] and residues 83–89 [6]. All predicted regions are consistent with the experimental ones (see figure 2(e)).

Reactive (or secondary) amyloidosis is characterized by the extracellular deposition of amyloid fibrils containing predominantly amyloid A protein (AA), which is a proteolytically derived fragment of serum amyloid A (SAA) protein. The N-terminus of AA protein (residues 1–11) was shown to be the amyloidogenic part of the molecule [55]. We predicted this region correctly (residues 1–6); however, two additional maxima (residues 16–20 and 66–70) appear on the packing density profile.

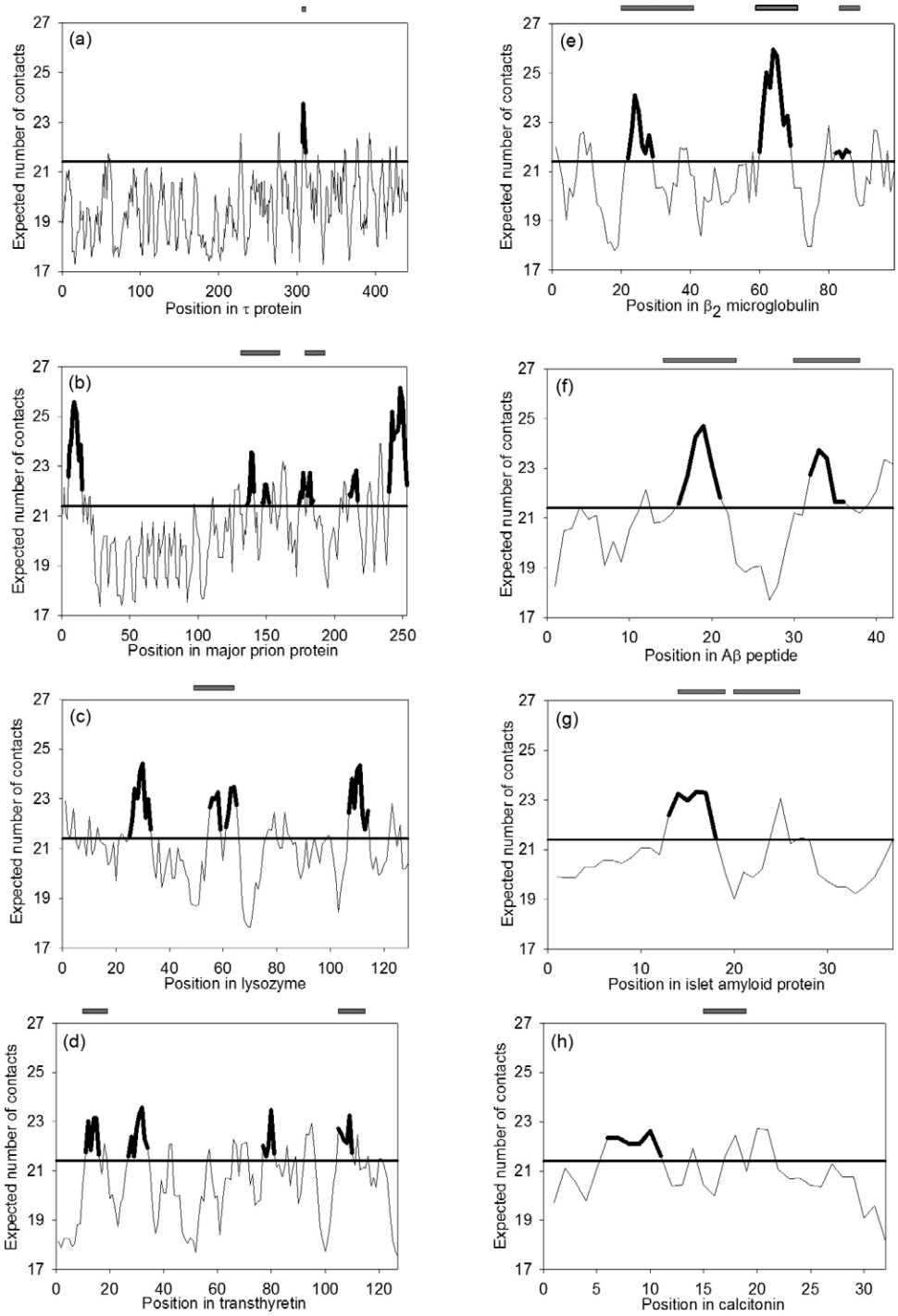

Figure 2. Contact profiles for the expected number of contacts for eight proteins and peptides from table 2. The predicted regions are drawn by thick lines; the experimentally localized amyloidogenic regions are drawn as grey rectangles above the profiles.

Medin is the main constituent of aortic medial amyloid. It is derived from a proteolytic fragment of lactadherin, a mammary epithelial cell-expressed glycoprotein that is secreted as part of the milk fat globule membrane. It was previously demonstrated that an octapeptide fragment of medin (residues 42–49, NFGSVQFV) forms typical well-ordered amyloid fibrils [56]. The last four residues (residues 47–50) have a large expected packing density, yet this region is not predicted by the rules of our algorithm (a region must be at least five residues).

It has been shown that residues 16–20 in amyloid β (Aβ) peptide are essential for peptide polymerization [57]. Also, solid-state NMR and site-directed spin labelling experiments suggest that residues 30–38 [58] form a β-strand in the fibrils. Our predictions (residues 16–21 and 32–36) are consistent with these experimental results (see figure 2(f)).

It has been shown that a fragment (residues 20–27) from amylin (also called human islet amyloid protein or hIAPP) is amyloidogenic and cytotoxic [59]. Other than this one, the shortest active fragments capable of self-assembly were found to be the pentapeptides FLVHS (residues 15–19) and NFLVH (residues 14–18) [60]. One of the fragments (residues 13–18) is correctly predicted by our method; however, the second amyloidogenic region (residues 20–27) has an expected packing density below the threshold (see figure 2(g)).

Alpha-synuclein is a major component of Lewy bodies in Parkinson's disease and is found to be associated with several other forms of dementia. The central fragment of α-synuclein (35 residues long), which has been isolated from purified amyloid of Alzheimer's disease brains [61], is called the non-Aβ-component of Alzheimer's disease amyloid (NAC). It has been shown that the N-terminal fragment of NAC (residues 3–18) forms aggregates and displays a transition from a random coil to a β-sheet structure [62]. On the contrary, the C-terminal fragment of NAC (residues 19–35) remains in solution with a random coil conformation under the same conditions [62]. No regions with an expected packing density over 20.4 are observed. The predicted region (residues 9–13) appears only if the threshold is 20.3. Thus, we consider this prediction to be a failure.

It has been shown that a peptide consisting of residues 15–19 of the human hormone calcitonin forms highly ordered fibrils, which are similar to those formed by the entire hormone sequence [63]. The profile for this peptide includes two regions with strong expected packing density: the first region corresponds to region 6–11, while the second one is separated by one residue with a low expected packing density. Thus, the second region is not predicted by the rules of our algorithm (see figure 2(h)).

Our predicted regions are consistent with known disease-related regions for 9 of 12 well-studied experimentally amyloidogenic peptides and proteins (transthyretin, β_2-microglobulin, lysozyme, prion protein and others). This result strongly indicates that the aggregation capability of a protein chain is one of the common properties of amyloid fibrils. Moreover, it should be noted that regions with a high packing density are often surrounded by amino acids that disrupt their amyloidogenic capability, regions with a weak expected packing density—probable amyloid breakers (see figure 2).

Here we also tested the ability of two other scales, hydrophobicity [39] and β-sheet propensity [40], to predict amyloidogenic regions and compared these results with our method of expected packing density. The thresholds for predictions were also obtained in a similar way (see figure 1). On the one hand, from 18 experimentally determined amyloidogenic regions, the expected packing density scale finds 14 regions (see tables 2 and 3), while the hydrophobicity scale finds nine and the β-sheet propensity scale finds four regions (in other words, the packing density scale misses four amyloidogenic fragments while the hydrophobicity scale misses nine fragments and β-sheet propensity scale misses 14). On the other hand, the scale of expected packing density finds 15 additional regions while the scale of hydrophobicity finds 17 extra

Table 3. Comparison of prediction of amyloidogenic regions using different scales.

	Scales			
Regions	Packing density	Hydrophobicity	β-sheet propensity	Hybrid
Predicted and confirmed by experiment	14	9	4	6
Predicted but not confirmed by experiment	15	17	10	6
Not predicted but observed in experiment	4	9	14	12

regions and the scale of β-sheet propensity finds 10 additional regions, the amyloidogenic role of which is not confirmed by experiment. The hybrid scale obtained by summation of all three scales (packing density, hydrophobicity and β-sheet propensity) gives worse numerical results (six fragments are correctly identified while there are six probable false positives) in spite of the fact that it is the best one for discriminating between fibril formers and fibril non-formers (see figure 1). Therefore, here we suggest a new property of peptides and proteins which form amyloid fibrils: regions with a strong expected packing density.

3.4. Searching for natively disordered regions

To test the quality of our predictions of natively unfolded regions in proteins, we have used two databases, of which one has 427 intrinsically disordered proteins and regions [38] and the other has 559 fully folded proteins [33]. The ROC curves obtained with differently sized sliding windows are shown in figure 3. The best result corresponds to the case when we construct the packing density profile smoothed over the sliding window of 41 residues; we chose 20.4 (the corresponding point is marked as a large circle) as the threshold (true positives 0.74 and false positives 0.03).

To test the quality of predictions obtained by our method compared to other methods of predicting disordered regions such as IUPred [33], DISOPRED2 [32], PONDR VL3H [28] and GlobPlot [31], we examined the same proteins that were used by Dosztanyi *et al* [33], who compared the quality of predictions obtained by their method (IUPred) with DISOPRED, PONDR VL3H and GlobPlot (the data on these methods were taken from [33]). These were a dataset of globular proteins (559 proteins) and long disordered protein segments (129 proteins). Table 4 demonstrates that our method (FoldUnfold) showed improved performance over these widely used methods (the averaging for our method is done in the same two ways as for the other methods [33]—over amino acid residues and over proteins).

4. Discussion

We demonstrate that expected packing density is a useful value for predicting both natively unfolded and amyloidogenic regions of a protein based only on its sequence. In figure 4, a distribution of the average packing densities of globular proteins is presented. The determined thresholds (21.4 for amyloidogenic regions and 20.4 for natively unfolded ones) correspond to the ends of this distribution.

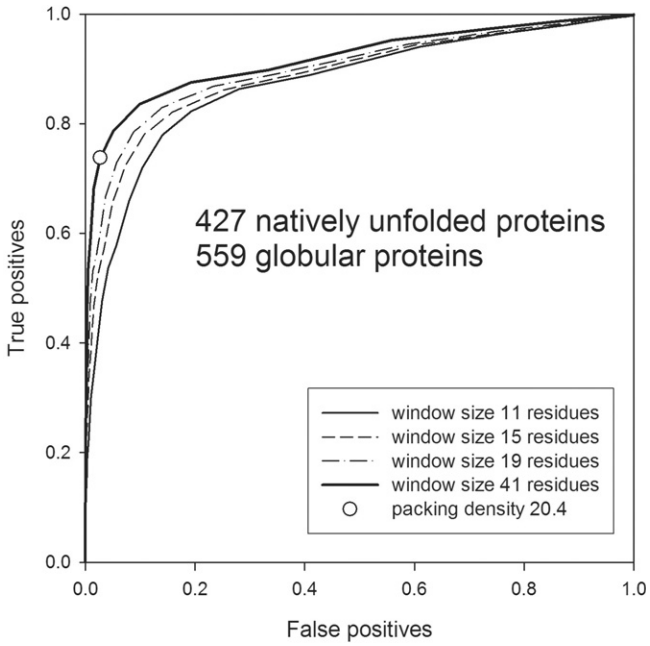

Figure 3. The ROC curves for prediction of natively unfolded regions. Each ROC curve corresponds to predictions with a specified (on the legend) size of the sliding window. The open circle corresponds to the value of packing density (20.4) that is chosen as a threshold.

Table 4. Performance of disorder prediction methods on datasets of globular proteins (559 proteins) and long disordered protein segments (129 proteins) [33].

	True positive rate: averaging is done over		False positive rate: averaging is done over	
Method	Residues	Proteins	Residues	Proteins
FoldUnfold (our method) [35]	0.851	0.716	0.051	0.076
IUPred [33]	0.763	0.679	0.053	0.055
PONDRVL3H [28]	0.663	0.607	0.050	0.078
DISOPRED2 [32]	0.664	0.491	0.050	0.069
GlobPlot [31]	0.330	0.304	0.181	0.197

Structures of peptides such as NNQQNY (derived from Sup35 protein [11]), KFFEAAAKKFFE (a designed 12-mer peptide [10]) and YTIAALLSPYS (derived from transthyretin [64]) confirm that the peptides adopt an extended β-strand conformation in amyloid fibrils. These fibrils achieve their stability through optimal values of main-chain and dihedral angles, as well as through extensive hydrophobic packing of side chains (hydrophobic template, Serrano's pattern—STVIIE) and salt bridge formation from polar side chains (polar template, Eisenberg's pattern—NNQQNY). It should be emphasized that between these two templates there probably exist many different intermediate variants. Our approach finds amyloidogenic regions closer to the hydrophobic template than to the polar one.

If amyloid fibril formation is a generic feature of proteins [3], some common properties of amino acid sequences possessing amyloidogenic propensities should be observed. Experimental data as well as theoretical analyses can help reveal the common structural and chemical properties for this process, one of which is the tight packing density.

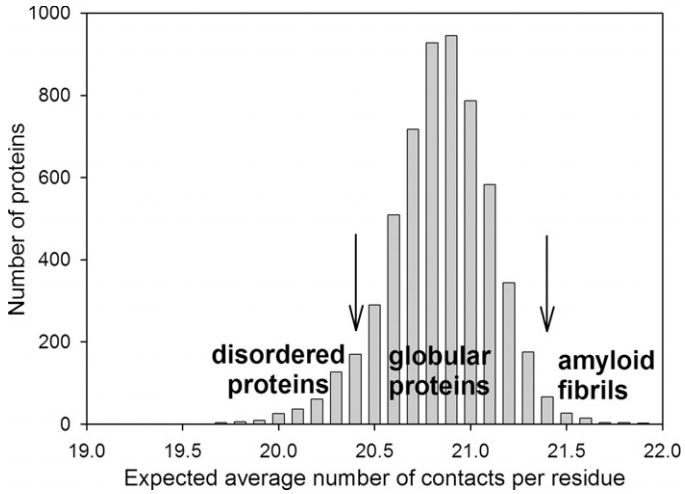

Figure 4. Histogram representing the distribution of 5829 globular protein domains as a function of the expected packing density. Arrows indicate upper and lower thresholds obtained from the ROC curves (see figures 1 and 3) which correspond to unusually strong and unusually weak expected packing densities.

We tried to collect all known amyloidogenic proteins and peptides for which disease-related regions are experimentally localized. By analysis of primary structure alone, we have demonstrated that regions that possess a strong expected packing density can be responsible for the amyloidogenic properties of a protein, while regions with a weak expected packing density correspond to disordered regions. A new concept is proposed that could aid in the understanding of protein folding, misfolding and amyloidosis.

Our study provides new insights into the process of amyloid formation. The results help to explain that the nature of the amyloidogenic propensity of proteins is linked to amino acid sequences with a high competence to form a large packing density. Our results can help determine the amyloidogenic propensity for amyloidogenic proteins for which the position of amyloidogenic regions now remains unexplored experimentally.

Acknowledgments

The authors are grateful to D Reifsnyder for help in producing the manuscript. This work was supported by the program 'Molecular and Cell Biology' of the Russian Academy of Sciences, by the Russian Foundation for Basic Research (grant 05-04-48750), by the Howard Hughes Medical Institute (no. 55005607) and by INTAS grant no. 05-1000004-7747.

References

[1] Jimenez J L, Guijarro J I, Orlova E, Zurdo J, Dobson C M, Sunde M and Saibil H R 1999 *EMBO J.* **18** 815–21
[2] Guijarro J I, Sunde M, Jones J A, Campbell I D and Dobson C M 1998 *Proc. Natl Acad. Sci. USA* **95** 4224–8
[3] Chiti F, Webster P, Taddei N, Clark A, Stefani M, Ramponi G and Dobson C M 1999 *Proc. Natl Acad. Sci. USA* **96** 3590–4
[4] Fandrich M, Fletcher M A and Dobson C M 2001 *Nature* **410** 165–6
[5] Bucciantini M, Calloni G, Chiti F, Formigli L, Nosi D, Dobson C M and Stefani M 2004 *J. Biol. Chem.* **279** 31374–82
[6] Ivanova M I, Sawaya M R, Gingery M, Attinger A and Eisenberg D 2004 *Proc. Natl Acad. Sci. USA* **101** 10584–9

[7] Yamamoto S, Hasegawa K, Yamaguchi I, Tsutsumi S, Kardos J, Goto Y, Gejyo F and Naiki H 2004 *Biochemistry* **43** 11075–82
[8] Pedersen J S, Christensen G and Otzen D E 2004 *J. Mol. Biol.* **341** 575–88
[9] Uversky V N, Gillespie J R and Fink A L 2000 *Proteins: Struct. Funct. Genet.* **41** 415–27
[10] Makin O S, Atkins E, Sikorski P, Johansson J and Serpell L C 2005 *Proc. Natl Acad. Sci. USA* **102** 315–20
[11] Nelson R, Sawaya M R, Balbirnie M, Madsen A O, Riekel C, Grothe R and Eisenberg D 2005 *Nature* **435** 747–9
[12] Fernandez A, Kardos J, Scott L R, Goto Y and Berry R S 2003 *Proc. Natl Acad. Sci. USA* **100** 6446–51
[13] Yoon S and Welsh W J 2004 *Protein Sci.* **13** 2149–60
[14] Lopez de la Paz M and Serrano L 2004 *Proc. Natl Acad. Sci. USA* **101** 87–92
[15] Esteras-Chopo A, Serrano L and Lopez de la Paz M 2005 *Proc. Natl Acad. Sci. USA* **102** 16672–7
[16] Thompson M J, Sievers S A, Karanicolas J, Ivanova M I, Baker D and Eisenberg D 2006 *Proc. Natl Acad. Sci. USA* **103** 4074–8
[17] Wright P E and Dyson H J 1999 *J. Mol. Biol.* **293** 321–31
[18] Dyson H J and Wright P E 2002 *Adv. Protein Chem.* **62** 311–40
[19] Wang J, Lu Q and Lu H P 2006 *PLOS Comp. Biol.* **2** e78
[20] Wang J, Zhang K, Lu H and Wang E 2006 *Phys. Rev. Lett.* **96** 168101
[21] Dyson H J and Wright P E 2005 *Nat. Rev. Mol. Cell. Biol.* **6** 197–208
[22] Wootton J C 1994 *Comput. Chem.* **18** 269–85
[23] Dunker A K, Garner E, Guilliot S, Romero P, Albrecht K, Hart J, Obradovic Z, Kissinger C and Villafranca J E 1998 *Pac. Symp. Biocomput.* 473–84
[24] Romero P, Obradovic Z, Kissinger C R, Villafranca J E, Garner E, Guilliot S and Dunker A K 1998 *Pac. Symp. Biocomput.* 437–48
[25] Romero P, Obradovic Z and Dunker A K 1999 *FEBS Lett.* **462** 363–7
[26] Galzitskaya O V, Surin A K and Nakamura H 2000 *Protein Sci.* **9** 580–6
[27] Vucetic S, Brown C J, Dunker A K and Obradovic Z 2003 *Proteins: Struct. Funct. Genet.* **52** 573–84
[28] Obradovic Z, Peng K, Vucetic S, Radivojac P, Brown C J and Dunker A K 2003 *Proteins: Struct. Funct. Genet.* **53** 566–72
[29] Radivojac P, Obradovic Z, Smith D K, Zhu G, Vucetic S, Brown C J, Lawson J D and Dunker A K 2004 *Protein Sci.* **13** 71–80
[30] Uversky V N 2002 *Eur. J. Biochem.* **269** 2–12
[31] Linding R, Jensen L J, Diella F, Bork P, Gibson T J and Russell R B 2003 *Structure* **11** 1453–9
[32] Ward J J, McGuffin L J, Bryson K, Buxton B F and Jones D T 2004 *Bioinformatics* **20** 2138–9
[33] Dosztanyi Z, Csizmok V, Tompa P and Simon I 2005 *J. Mol. Biol.* **347** 827–39
[34] Garbuzynskiy S O, Lobanov M Yu and Galzitskaya O V 2004 *Protein Sci.* **13** 2871–7
[35] Galzitskaya O V, Garbuzynskiy S O and Lobanov M Yu 2006 *Mol. Biol. (Moscow)* **40** 341–8
[36] Murzin A G, Brenner S E, Hubbard T and Chothia C 1995 *J. Mol. Biol.* **247** 536–40
[37] Bairoch A and Apweiler R 2000 *Nucleic Acids Res.* **28** 45–8
[38] Vucetic S, Obradovic Z, Vacic V, Radivojac P, Peng K, Iakoucheva L M, Cortese M S, Lawson J D, Brown C J, Sikes J G, Newton C D and Dunker A K 2005 *Bioinformatics* **21** 137–40
[39] Fauchere I I and Pliska V 1983 *Eur. J. Med. Chem-Chim. Ther.* **18** 369–75
[40] Minor D L Jr and Kim P S 1994 *Nature* **371** 264–7
[41] Rudall K M 1952 *Adv. Protein Chem.* **7** 253–90
[42] Von Bergen M, Friedhoff P, Biernat J, Heberle J, Mandelkow E M and Mandelkow E 2000 *Proc. Natl Acad. Sci. USA* **97** 5129–34
[43] Morrissey M P and Shakhnovich E I 1999 *Proc. Natl Acad. Sci. USA* **96** 11293–8
[44] Speare J O, Rush T S III, Bloom M E and Caughey B 2003 *J. Biol. Chem.* **278** 12522–9
[45] Thompson A, White A R, McLean C, Masters C L, Cappai R and Barrow C J 2000 *J. Neurosci. Res.* **62** 293–301
[46] Torrent J, Alvarez-Martinez M T, Liautard J P, Balny C and Lange R 2005 *Protein Sci.* **14** 956–67
[47] Hamidi A K, Liepnieks J J, Nakamura M, Parker F and Benson M D 1999 *Biochem. Biophys. Res. Commun.* **257** 584–8
[48] Krebs M R, Wilkins D K, Chung E W, Pitkeathly M C, Chamberlain A K, Zurdo J, Robinson C V and Dobson C M 2000 *J. Mol. Biol.* **300** 541–9
[49] Frare E, Polverino de Laureto P, Zurdo J, Dobson C M and Fontana A 2004 *J. Mol. Biol.* **340** 1153–65
[50] Chamberlain A K, MacPhee C E, Zurdo J, Morozova-Roche L A, Hill H A, Dobson C M and Davis J J 2000 *Biophys. J.* **79** 3282–93
[51] MacPhee C E and Dobson C M 2000 *J. Mol. Biol.* **297** 1203–15
[52] Jaroniec C P, MacPhee C E, Astrof N S, Dobson C M and Griffin R G 2002 *Proc. Natl Acad. Sci. USA* **99** 16748–53

[53] Kozhukh G V, Hagihara Y, Kawakami T, Hasegawa K, Naiki H and Goto Y 2002 *J. Biol. Chem.* **277** 1310–5
[54] Jones S, Manning J, Kad N M and Radford S E 2003 *J. Mol. Biol.* **325** 249–57
[55] Patel H, Bramall J, Waters H, De Beer M C and Woo P 1996 *Biochem. J.* **318** 1041–9
[56] Reches M and Gazit E 2004 *Amyloid* **11** 81–9
[57] Tjernberg L O, Callaway D J, Tjernberg A, Hahne S, Lilliehook C, Terenius L, Thyberg J and Nordstedt C 1999 *J. Biol. Chem.* **274** 12619–25
[58] Torok M, Milton S, Kayed R, Wu P, McIntire T, Glabe C G and Langen R 2002 *J. Biol. Chem.* **277** 40810–5
[59] Azriel R and Gazit E 2001 *J. Biol. Chem.* **276** 34156–61
[60] Mazor Y, Gilead S, Benhar I and Gazit E 2002 *J. Mol. Biol.* **322** 1013–24
[61] Ueda K, Fukushima H, Masliah E, Xia Y, Iwai A, Yoshimoto M, Otero D A, Kondo J, Ihara Y and Saitoh T 1993 *Proc. Natl Acad. Sci. USA* **90** 11282–6
[62] Bodles A M and Irvine G B 2004 *Protein Pept. Lett.* **11** 271–9
[63] Haspel N, Zanuy D, Ma B, Wolfson H and Nussinov R 2005 *J. Mol. Biol.* **345** 1213–27
[64] Jaroniec C P, MacPhee C E, Bajaj V S, McMahon M T, Dobson C M and Griffin R G 2004 *Proc. Natl Acad. Sci. USA* **101** 711–6