Supplementary Information for

Metastability of Native Proteins and the Phenomenon of Amyloid Formation

Andrew J. Baldwin[1], Tuomas P. J. Knowles[1], Gian Gaetano Tartaglia[1], Anthony W. Fitzpatrick[1], Glyn L. Devlin[1], Sarah Shammas[1], Christopher A. Waudby[1], Maria F. Mossuto[1], Sarah Meehan [1], Sally L. Gras[2], John Christodoulou[3], Spencer J. Anthony-Cahill[4], Paul D. Barker[1], Michele Vendruscolo[1] and Christopher M. Dobson[1]

[1] Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, UK

[2] Bio21 Molecular Science and Biotechnology Institute and Department of Chemical and Biomolecular Engineering, University of Melbourne, Parkville, VIC 3010, Australia

[3] Research Department of Structural and Molecular Biology, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK

[4] Department of Chemistry, Western Washington University Bellingham, WA 98225-9150, USA

S1. Methods and materials

Fibril preparation

All fibrils used in this study were prepared in two stages. Protein samples were first transfered directly into the buffer appropriate for fibril formation (see main text) and incubated at room temperature, resulting in spontaneous aggregation. The kinetics of the aggregation reaction were followed by monitoring the change in fluorescence of the dye thioflavin-T on addition to an aliquot of the reaction mixture [1]. The aggregation mixture was deemed to have reached its conclusion when the observed fluorescence signal was found to remain constant for a period of 24 hours. The total time required for this depended on the exact system under study and the ambient solution conditions, typically on the order of several days. The aggregated material was then separated from un-aggregated material by ultracentrifugation in a Beckman Optima TLX ultracentrifuge at 90,000 rpm for 40 min, as described in the text, and the pellet resuspended. The aggregation reaction was then repeated as described above, with the addition of 5% molar ratio of the previously purified pellet as seed. The resulting purified material was assayed by electron microscopy [1] to examine the morphology of the aggregates. In cases where less than ca. 90% of the material was present as fibrils, this procedure was repeated, using the most recently purified batch of aggregated material as seed for the subsequent aggregation reaction. In the case of α-synuclein aggregation, up to four rounds of consecutive aggregation reactions were necessary to produce a homogeneous suspension of fibrils, whereas for insulin, only one round was required. It is likely that this method selects for stable fibrillar aggregates over relatively unstructured aggregates that might otherwise be kinetically favoured. When fibrils were produced according to this protocol, the denaturation curves and hence fibril stabilities were found to be highly reproducible.

Denaturation assay

Homogeneous fibril suspensions were added at a final concentration of 5-70 μM to solutions containing a specified concentrations of either guanidinium thiocyanate (GdSCN) or guanidinium hydrochloride (GdHCl) in the same buffer used to produce aggregated protein. High concentrations of GdHCl were unable to appreciably dissociate the majority of the fibrils tested, necessitating the use of the stronger denaturant, GdSCN. Solutions were thoroughly mixed by vortexing and incubated at room temperature for 24 hours, prior to centrifugation in a Beckman Optima TLX ultracentrifuge at 90,000 rpm for 40 min to separate un-aggregated and aggregated protein, as described in the text. This incubation time was experimentally verified in all cases to be sufficient for the samples to reach equilibrium.

When GdHCl was used, the protein concentration in the supernatant was quantified by measuring the absorbance at 280 nm. Due to the absorbance of the GdSCN in the UV frequency range, absorption at 280 nm could not be used directly to determine the concentration of soluble protein when this denaturant was used. Instead the supernatants were first diluted 30 fold in distilled water, then mixed in a 1:1 volumetric ratio with Bradford reagent (Sigma-Aldrich Co. Ltd., Dorset, UK). The resultant change in absorbance at 595 nm was used to quantify the quantity of soluble protein, M_S.

Tests for equilibrium

In reaching equilibrium a system evolves to a constant composition from an arbitrary initial condition. We performed several tests on the aggregates formed from each protein and peptide system to demonstrate that this was the case: 1) We took a solution of fibrils that have been almost com-

pletely dissociated by denaturant (M_S/M_T 0.8) and diluted the denaturant to a value where we expect solution conditions that once again favour fibrils (M_S/M_T 0.2). Our finding was that over the course of several days, the fibrils reform, and the solution relaxes to its expected composition; 2) At a given solution condition, additional monomeric material was added and the solution was left for several days. The ratio M_S/M_T was then re-measured. In all cases it was found to have returned to the equilibrium level; 3) A fibril solution, in the absence of free monomers can be prepared by ultracentrifugation, as described above. In the cases of TTR and SH3 where the quantity of free monomer in the absence of denaturant at equilibrium is relatively high. Consequently, it is possible to follow the re-emergence of monomeric material over time, as the fibrils dissociate. The systems were again found to relax towards the expected equilibrium monomer:fibril composition.

Comparison of GdHCl and GdSCN stability measurements

Estimates of free energy obtained from guanadinium thiocyanate and guanadinium hydrochloride were compared directly in this work. Empirically, an adjustment of $[D] = \alpha[D]_{\mathrm{GdSCN}}/([D]_{\mathrm{GdSCN}} + \alpha)$ where $\alpha = 6.47$ has been proposed [2] when performing experiments to compare the stability of globular proteins obtained using GdSCN as the denaturant, to those obtained using GdHCl. When we employ such an adjustment in this work, the resulting free elongation energy values, ΔG_{adj} are highly correlated with those that are not adjusted, such that $\Delta G_{\mathrm{adj}} = 2.00\Delta G_0 + 25.2$. The Pearson's correlation coefficient, r^2, between ΔG_{adj} and ΔG_0 was 0.98. Due to such a strong linear correlation between the two energy measurements, neither the conclusions nor scaling relations described in the main text of the paper are effected by such an adjustment. It is unclear whether such an adjustment is appropriate in this instance, as typically significantly higher concentrations of GdSCN are required for 50% denaturation of fibrils when compared to globular proteins. The free energies we report in the the main text are therefore not adjusted.

S2. Linear polymerisation model

In the spirit of the Oosawa's linear polymerisation model [3], we considered the scheme $M + F_{i-i} \rightleftarrows F_i$, in which the populations of aggregates F_i of size i are in equilibrium with the population of monomers, $M_S = F_1$. In this case the equilibrium constant is $K_{eq} = \frac{[F_i]}{[M][F_{i-1}]}$ and the total concentration is $M_T = \sum_{i=1}^{\infty} i[F_i] = \sum_{i=1}^{\infty} iK_{eq}^{i-1}M_S^i = \frac{M_S}{(1-K_{eq}M_S)^2}$, where the last step requires use of the series for $(1-x)^{-2}$. By rearranging this expression we obtain $K_{eq} = \frac{1}{M_S} - \frac{1}{\sqrt{M_S M_T}}$. At equilibrium therefore $M_S = \frac{1}{K_{eq}} - \frac{1-\sqrt{1+4K_{eq}M_T}}{2M_T K_{eq}^2}$. In the limit $M_S > M_T$, the equilibrium expression reduces to the situation where fibril ends E are in equilibrium with monomers $M + E \rightleftarrows E$, with the equilibrium constant $K_{eq} = 1/M_S^{\max}$.

S3. Sub-linear scaling behaviour of the stability of amyloid fibrils.

In order to provide an explanation for such sub-linear scaling behaviour, we considered how geometric and topological constraints, both of which increase in significance with growing chain lengths, limit the independent optimisation of the large number of individual interactions characteristic of polypeptide chains in amyloid fibrils. We started from the observation that a general characteristic of polypeptide systems in their amyloid form is that their energetics are defined by a large number of individual interactions, ϵ_i, which average to a total energy $\sum_i \epsilon_i \ll \sum_i |\epsilon_i|$ that will be less than the sum of the absolute values of the individual interactions. This feature can be considered within an approach analogous to the random energy model (REM) derived for spin glasses[4] and polymers[5]. Indeed, according to the central limit theorem, the

probability distribution for the free energy G of a polypeptide chain in a fibril with $N \gg 1$ follows an approximately Gaussian distribution given by $P(G) = \frac{1}{\sigma_\epsilon \sqrt{2\pi z N}} \exp\left(-\frac{(G - Nz\langle\epsilon\rangle)2}{2Nz\sigma_\epsilon}\right)$, where z is the average number of contacts per residue, and σ_ϵ and $\langle\epsilon\rangle$ are the standard deviation and mean, respectively, of the (unknown) probability distribution characterising the free energy of individual contacts. A large majority of states will then have a free energy higher than $\Delta G^0_{\rm el} = Nz\langle\epsilon\rangle - \sigma_\epsilon\sqrt{Nz}$, which therefore represents an estimate of the lowest free energy state that the system typically occupies. From this approach, we obtain an expression for the energy for each residue $\frac{\Delta G^0_{\rm el}}{N} = z\langle\epsilon\rangle - \sigma_\epsilon\sqrt{z}\,N^{-1/2}$, which corresponds to an estimate of a value of -1/2 for the exponent γ.

It is also possible to rationalise the origin of the sub-linear scaling behaviour discussed in this work using a geometric argument, rather than an energetic one as above, by considering the interactions of a polypeptide chain with its nearest neighbours in a fibril, and the corresponding contribution to the free energy difference between a chain free in solution and assembled within a fibril. In addition to interactions that scale extensively with the number of residues, further interchain interactions are expected to scale with the contact area of a given chain that is exposed to the neighbouring chains, and so $\Delta G^0_{\rm el} \propto \epsilon_V R^3 + \epsilon_S R^2$, where R is a measure of the linear dimension of a chain when packed into a fibril. Dimensionality then requires that $\Delta G^0_{\rm el} = \epsilon_V N + \epsilon_S N^{2/3}$, and hence the energy change per residue is given by $\frac{\Delta G^0_{\rm el}}{N} = \epsilon_V + \epsilon_S\, N^{-1/3}$, also resulting in a sub-linear power law for the scaling of the free energy with residue number, with a value of -1/3 for the exponent γ. The agreement between the experimental data and these power laws is excellent (Figure 1C in the main text), and indeed the average deviation from the calculated free energies is less than 20%.

S4. Optimal fibril length

The sequence length expected to give maximally stable fibrils, N_{opt}, is given by:

$$\begin{aligned} \Delta G_{\rm el} &= \epsilon_0 N + \epsilon_1 N^{\gamma+1} \\ \frac{\delta \Delta G_{\rm el}}{\delta N} &= \epsilon_0 + (\gamma+1)\epsilon_1 N^{\gamma} \\ N_{opt} &= \left(\frac{-\epsilon_0}{\epsilon_1(\gamma+1)}\right)^{\frac{1}{\gamma}} \end{aligned} \tag{1}$$

S5. Native stabilities

S6. Amino acid sequences

Protein	Species	Chain length	ΔG / kJ mol^{-1}
Abp1 SH3	Common Yeast	69	-13
ACBP	Bovine	88	-23.7
ACBP	Rat	87	-25.3
ACBP	Common Yeast	87	-32.6
ADAh2	Human	81	-17.1
Apo-azurin	Pseudomonas aeruginosa	129	-29.2
CheW	Thermotoga maritima	152	-32.5
CI2	Barley	66	-32.5
CI2	Barley	66	-31.8
CI2	Barley	66	-31.8
CTL9	Bacillus stearothermophilus	93	-27.2
EC0298	E. coli	89	-11.4
FKBP12	Human	110	-23.4
FKBP12	Human	110	-23.1
FRB	Human	100	-26.2
FRB	Human	100	-36.9
FRB	Human	100	-17.1
FRB	Human	100	-26.5
FRB	Human	108	-29.4
FRB	Human	108	-32.3
FynSH3	Human	89	-20.3
GW1	Listeria monocytogenes	86	-15.5
HPr	E. coli	85	-20.8
Im7	E. coli	87	-26.5
Im7*	E. coli	96	-11.7
Im9	E. coli	94	-20.9
Im9	E. coli	86	-26.1
L23	Thermus thermophilus	97	-21.2
Lambda Repressor	lambda phage	82	-21.2
mAcP	Horse	100	-20.4
NTL9	Bacillus stearothermophilus	56	-17.3
Protein L	Peptostreptococcus magnus	72	-19.9
raf RBD	Human	81	-26
S6	Thermus thermophilus	101	-34.7
Sho1 SH3	Unknown	77	-9.4
Spectrin SH3	Chicken	64	-13.9
SrcSH2	Rous sarcoma virus	112	-31
SrcSH3	Chicken	64	-15.9
Tm1083	Bovine	127	-38.2
U1A	Human	103	-34.8
Ubiquitin	Human	78	-34.2
Urm1	Common Yeast	102	-13
VlsE	Lyme disease spirochete	339	-23.7

Table S1: Native stabilities from the Protein Folding Database (PFD) [6].

Protein	Species	Chain length	ΔG / kJ mol^{-1}
lamba repressor	lambda phage	80	-12.6
lambda repressor G46AG48A	lambda phage	80	-20.1
ACBP	bovine	86	-29.7
ACBP	rat	86	-25.5
ACBP	yeast	86	-32.6
CytochromeC		104	-28.9
Cytochrome C (ox)	Horse	104	-74.1
CytochromeC (ox)	Yeast	103	-61.1
CspB	Bacillus subtilis	67	-12.6
CspB	Bacillus subtilis	67	-11.3
CspB	Bacillus caldolyticus	66	-20.1
CspB	Thermotoga maritima	68	-26.4
CspA		69	-12.6
CspA		69	-12.1
a-spectrin		62	-12.1
Src		64	-17.2
PI3 kinase		84	-14.2
Fyn		67	-25.1
FNIII		90	-5.0
TWIg18		93	-16.3
TNfn3 (short form)		90	-20.1
TNfn3 (long form)		90	-28.0
CD2		98	-33.5
CI2 activaton domain		64	-29.3
Procarboxypeptidase A2 (ADAh2)		81	-17.2
Arc repressor (single chain)		106	-26.4
Ubiquitin V26A		76	-16.3
Ubiquitin V26G		76	-15.5
IgG binding domain of protein L		62	-19.2
Splicosomal protein U1A	streptococcus	102	-38.9
Hpr		85	-19.2
FKBP12		107	-23.0
Muscle AcP		98	-22.6

Table S2: Native stabilities from [7].

$TTR_{(105-115)}$

YTIAALLSPY S

TTRRGD

YTIAALLSPY SGGRGDS

SH3

GSMSAEGYQY RALYDYKKER EEDIDLHLGD ILTVNKGSLV ALGFSDGQEA KPEEIGWLNG YNETTGERGD
FPGTYVEYIG RKKISP

Human Lysozyme

KVFERCELAR TLKRLGMDGY RGISLANWMC LAKWESGYNT RATNYNAGDR STDYGIFQIN SRYWCNHGKT
PGAVNACHLS CSALLQDNIA DAVACAKRVV RDPQGIRAWV AWRNRCQNRD VRQYVQGCGV

Disulphide linkages between 6-128, 30-116, 65-81, 77-95

Insulin

A chain -

GIVEQCCASV CSLYQLENYC N

B chain -

FVNQHLCGSH LVEALYLVCG ERGFFYTPKA

Disulphide linkages between A6-A11, A7-B7,A20-B19

αB-crystallin

DIAIHHPWIR RPFFPFHSPS RLFDQFFGEH LLESDLFPTS TSLSPFYLRP PSFLRAPSW FDTGLSEMRL
EKDRFSVNLD VKHFSPEELK VKVLGDVIEV HGKHEERQDE HGFISREFHR KYRIPADVDP LTITSSLSSD
GVLTVNGPRK QVSGPERTIP ITREEKPAVT AAPKK

α-synuclein

MDVFMKGLSK AKEGVVAAAE KTKQGVAEAA GKTKEGVLYV GSKTKEGVVH GVATVAEKTK EQVTNVGGAV
VTGVTAVAQK TVEGAGSIAA ATGFVKKDQL GKNEEGAPQE GILEDMPVDP DNEAYEMPSE EGYQDYEPEA

β2-microglobulin

MSRSVALAVL ALLSLSGLEA IQRTPKIQVY SRHPAENGKS NFLNCYVSGF HPSDIEVDLL
KNGERIEKVE HSDLSFSKDW SFYLLYYTEF TPTEKDEYAC RVNHVTLSQP KIVKWDRDM

Aβ(1-40)

DAEFRHDSGY EVHHQKLVFF AEDVGSNKGA IIGLMVGGVV

Table S3: Sequences of polypeptides and proteins studied

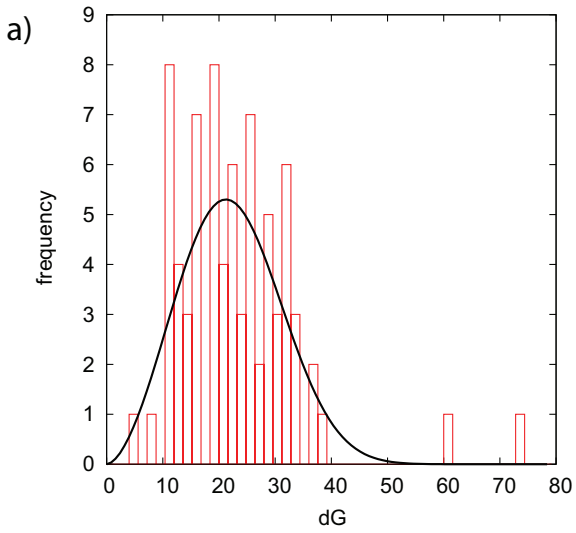

Figure S1: Histogram of the folding energies of the 76 natively folded proteins studied. The histogram is fitted to a Gaussian function of average position 21 kJ mol^{-1} and standard deviation of 9.25 kJ mol^{-1}. The two proteins with exceptionally high stabilities are cytochrome C derived from horse (-74 kJ mol^{-1}) and yeast (-61 kJ mol^{-1}).

References

[1] J. Zurdo, J. I. Guijarro, and C. M. Dobson. Preparation and characterization of purified amyloid fibrils. *J Am Chem Soc*, 123(33):8141–2, 2001.

[2] J. Clarke, E. Cota, S. B. Fowler, and S. J. Hamill. Folding studies of immunoglobulin-like beta-sandwich proteins suggest that they share a common folding pathway. *Structure*, 7(9):1145–53, 1999.

[3] F. Oosawa and M. Kasai. A theory of linear and helical aggregations of macromolecules. *J Mol Biol*, 4:10–21, 1962.

[4] Bernard Derrida. Random-energy model: An exactly solvable model of disordered systems. *Physical Review B*, 24(5):2613 LP – 2626, 1981.

[5] J. D. Bryngelson and P. G. Wolynes. Spin glasses and the statistical mechanics of protein folding. *Proc Natl Acad Sci U S A*, 84(21):7524–8, 1987.

[6] http://www.foldeomics.org/pfd/public_html/index.php.

[7] A. R. Fersht. *Structure and mechanism in protein science*. W.H. Freeman, New York, 1998.