Simultaneous quantification of protein order and disorder

Pietro Sormanni[1], Damiano Piovesan[2], Gabriella T Heller[1], Massimiliano Bonomi[1], Predrag Kukic[1], Carlo Camilloni[3], Monika Fuxreiter[4], Zsuzsanna Dosztanyi[5], Rohit V Pappu[6], M Madan Babu[7], Sonia Longhi[8], Peter Tompa[9], A Keith Dunker[10], Vladimir N Uversky[11], Silvio C E Tosatto[2] & Michele Vendruscolo[1]*

Nuclear magnetic resonance spectroscopy is transforming our views of proteins by revealing how their structures and dynamics are closely intertwined to underlie their functions and interactions. Compelling representations of proteins as statistical ensembles are uncovering the presence and biological relevance of conformationally heterogeneous states, thus gradually making it possible to go beyond the dichotomy between order and disorder through more quantitative descriptions that span the continuum between them.

The discovery of disordered proteins, which constitute about one-third of the human proteome and are crucial for regulation and signaling[1–3], has profoundly shaken the long-held paradigm that proteins fold into well-defined native structures whose atomic coordinates can be determined almost univocally. This finding has been followed by a polarization of the terms 'order' and 'disorder', which, in hindsight, has been largely prompted by a lack of techniques capable of fully characterizing the dynamics of proteins. Protein disorder was initially defined as 'absence of structure', for example, from regions of missing coordinates in native structures determined through X-ray crystallography[1–3]. Such a definition implies that order and disorder are mutually exclusive, while in fact protein structures and dynamics are closely related and central to the functions of these molecules.

In this Commentary, we discuss how the development of methods capable of simultaneously determining structure and dynamics of proteins, including, in particular, nuclear magnetic resonance (NMR) spectroscopy, is gradually making it possible to supersede this rather artificial polarization between order and disorder. We anticipate that the introduction of increasingly quantitative descriptions of structure and dynamics will provide compelling insights into the molecular mechanisms underlying protein behavior.

An order-disorder continuum

It is increasingly recognized that the native states of proteins range from fully ordered to almost completely disordered, with all intermediate situations in between[4,5] (**Fig. 1**). In this context, it is becoming generally accepted that the functional interpretation of structural data is often complicated by the fact that most current methodologies, being poorly equipped to describe motions, tend to determine only the most representative 'static' structure within the ensemble populated in solution. In this context, advances in kinetic protein crystallography[6,7] and integrative structural biology[6,8] (i.e., the combination of various complementary methods of structural determination) are providing atomic-resolution descriptions of the states sampled on the picosecond to nanosecond timescales (see **Box 1**). Approaches of this type have already shown that many proteins for which a tightly packed crystal structure determined at cryogenic temperatures is available are actually quite dynamic, particularly in regions important for function, interactions, and allosteric regulation[7]. There is, however, an even greater need to further develop techniques capable of also accurately describing the motions of larger amplitudes and longer timescales that are typical of disordered proteins (see **Box 1**). As progress is made in this direction, the dichotomy between protein order and disorder will gradually be replaced by quantitative descriptions of the range of situations between these two extremes.

The rise of NMR spectroscopy

One of the most spectacular recent developments in structural biology has been provided by NMR methods capable of quantitatively determining the structural fluctuations of proteins[4,9–13], offering powerful means to achieve a simultaneous characterization of order and disorder in proteins. These developments are firmly based on the long history of NMR spectroscopy. The initial success of this technique was due to its ability to determine in solution the structures of native states with a structural accuracy that in the best cases is comparable with that of X-ray crystallography in the solid state[14]. At variance with X-ray crystallography, however, NMR spectroscopy can also shed light on the dynamics of proteins in solution on a wide range of timescales (**Box 1**, top). Chemical shifts and residual dipolar couplings, which span up to the millisecond timescale, can probe processes ranging from ligand binding and allostery to catalysis and folding[10–12,15–18]. Other NMR measurements, such as those exploiting nuclear Overhauser effects (NOEs), and longitudinal (R1) and transverse (R2) relaxation processes, which can be used to probe the picosecond to nanosecond timescales, are informative of side chain rotations and local motions[6,9,12,14]. Paramagnetic relaxation enhancement (PRE) and electron paramagnetic resonance (EPR) experiments report on the dynamics on the microsecond timescale, typical of the formation of folding intermediates and of ligand-binding processes[10,12]. Extending into the millisecond timescale, NMR R1ρ rotating-frame relaxation and Carr–Purcell–Meiboom–Gill (CPMG) experiments provide information into folding and binding intermediates[10]. On the longest timescale reachable by NMR techniques, real-time NMR and hydrogen–deuterium (H–D) exchange data can probe dynamics beyond the second timescale, typical of the folding of complex proteins[10]. Furthermore, as NMR measurements report on average values over the structural fluctuations of proteins,

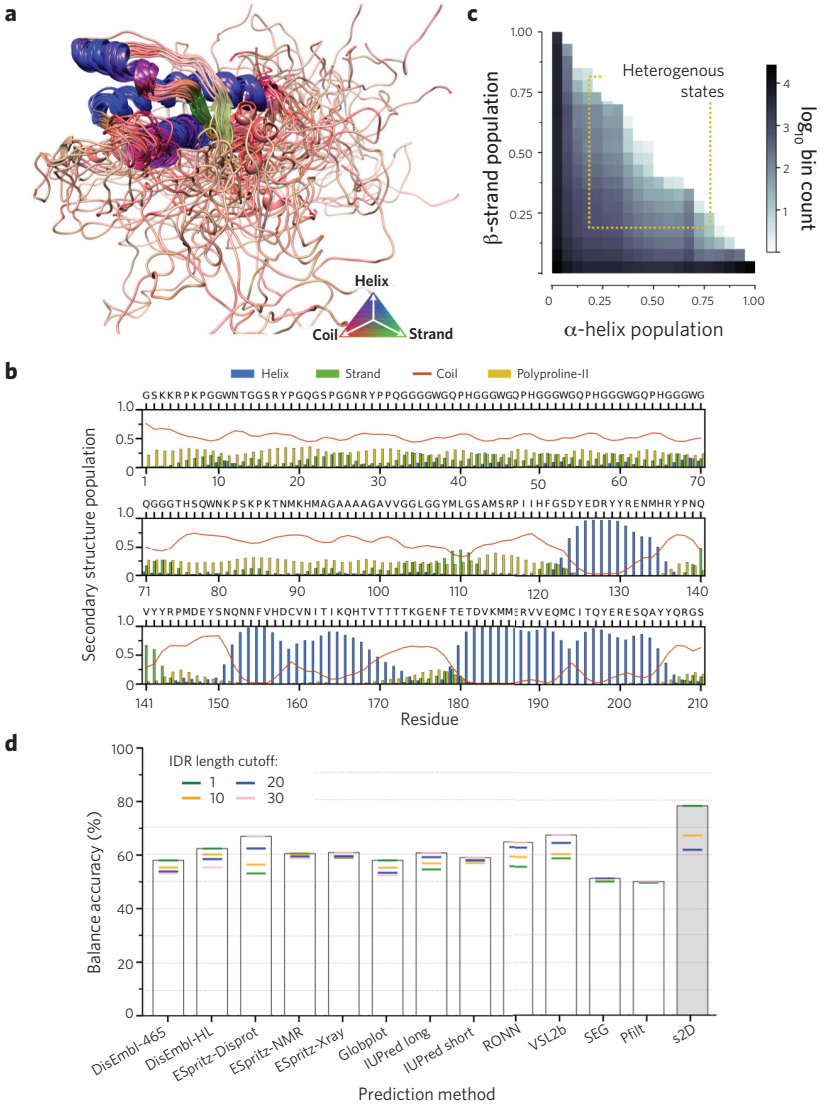


Figure 1 | Protein structure and dynamics can be represented effectively through structural ensembles. (**a**) Structural ensemble of the human prion protein calculated using the metainference method[22]. (**b**) Corresponding secondary structure populations as determined from NMR chemical shifts using the δ2D method[19] (BMRB ID 4402). (**c**) Scatter plot of the α-helix and β-strand populations for all residues in the PODD data set. The dashed rectangle highlights residues in heterogeneous regions, which populate more than one type of secondary structure element. (**d**) Bar plot of the balanced accuracy of commonly-used sequence-based methods of predicting disorder (*x*-axis) on a subset of the PODD data set corresponding to chemical shifts measured on monomeric proteins under physiological conditions. Regions are defined in this panel as disordered if they comprise at least *L* consecutive residues with a population of both α-helix and β-strand smaller than 0.5 (*L* = 1,10,20,30 as in the legend) or as ordered otherwise. The column for the s2D method[23] is in gray as some sequences in the data set are part of its training set.

they can provide information about the equilibrium dynamics of these molecules by enabling the determination of the different structures that they populate (i.e., their structural ensembles; see **Box 1**). In particular, NMR spectroscopy is playing a crucial role in providing structural information about states that are intrinsically highly dynamic and cannot be crystallized[10–13,19–21]. It is also becoming increasingly possible to use NMR spectroscopy to determine transition rates between different states, thus enabling the description of nonequilibrium dynamic processes[10] (see **Box 1**).

Structural ensemble challenges

Despite the optimism about the great potential of NMR spectroscopy to offer an accurate determination of protein structural ensembles, this task remains extremely difficult. In most cases, experimental data represent averages weighted over all populated states, which poses a 'deconvolution problem', as one has to resolve the different states that yield the measured averages. Furthermore, these averages provide sparse information—often coming from different types of experiments—concerning, for example, only certain bond angles and certain interatomic distances, which needs to be integrated coherently. Finally, experimental data are affected by random and systematic errors, and the energy functions employed in computer simulations are only approximations of the actual interactions between the atoms comprising proteins and solvent. Several techniques with varying degrees of sophistication have been developed to integrate multiple types of experimental data with *a priori* knowledge (for example, about force fields) to model structural ensembles[4,9,10,12,13,16,18,22]. The ensembles generated by applying these techniques have demonstrated the existence of different degrees of dynamics, ranging from functionally relevant small-scale native state fluctuations[9,15,17] to the large-amplitude motions in the conformationally heterogeneous states populated by disordered proteins[4,10–13,20,21].

Toward structural ensemble repositories

Quite generally, any method of ensemble modeling represents a compromise between the following factors: (1) the quality of the resulting structural ensemble, particularly in terms of the amount of information that can be extracted from it, (2) the amount and quality of available experimental data, and (3) the time and resources needed for its application. The Protein Data Bank (PDB) currently contains only a very small number of structural ensembles. While protein structures determined by NMR spectroscopy are often deposited as multiple models that individually fit the NMR data[9], they do not contain the statistical populations of the different states; thus, they are not 'statistical ensembles' but rather 'uncertainty ensembles'. To address this problem, the Protein Ensemble Database (PED)[13] has recently been compiled. However, its still relatively small number of entries (currently 24) reflects the fact that accurate structural ensemble calculations remain highly demanding in terms of both computational resources and quantity and quality of required experimental data. Moreover, many structural ensembles in the PED do not yet include information about statistical populations, making it hard to identify the most relevant states. Overall, although ensemble determination is still daunting, the availability of increasingly accurate experimental and theoretical methods as well as the rapid growth of computing power offer hope for the future development of a large repository of structural ensembles able to describe the properties of proteins in solution more comprehensively than static structures.

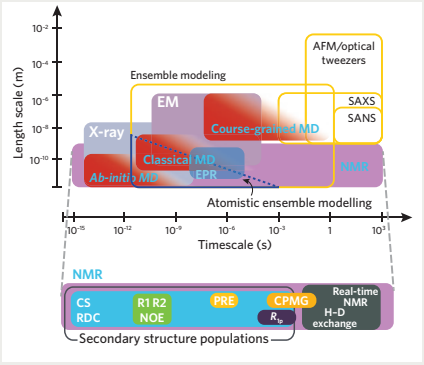
Two-dimensional ensembles

Given the challenges described above in determining structural ('three-dimensional') ensembles, a complementary strategy is to focus on 'two-dimensional ensembles', which are generally easier to calculate while still providing quantitative information about relevant properties of highly dynamic states of proteins. In this context, the Protein Order and Disorder Database (PODD, http://www-mvsoftware.ch.cam.ac.uk/index.php/podd) contains the two-dimensional ensembles, in terms of secondary structure populations, of about 5,000 proteins, determined directly from NMR chemical shifts using the δ2D method[19]. A structural ensemble of the human prion protein (**Fig. 1a**) and the corresponding secondary structure populations (**Fig. 1b**) are compared here as an example. While these two-dimensional ensembles do not provide the probability distributions of atomic coordinates or of tertiary contacts, they do offer useful estimates of local stability and structural heterogeneity. A large fraction of all residues cataloged in PODD are found in heterogeneous regions of proteins that populate both α-helices and β-strands (**Fig. 1c**). The main advantages of using secondary structure populations is that their determination is computationally inexpensive, and backbone chemical shifts are often readily measurable. Furthermore, when chemical shifts are not available, secondary structure populations can be predicted from amino acid sequences, for instance, using the s2D method[23].

The structural characterization of proteins in PODD provides an illustration of the concept of continuum between order and disorder. Any separation between them is not absolute, but depends on the introduction of an arbitrary cut-off value on the populations to break the continuity between them. To verify that the most dynamic regions present in PODD, in which populations are derived from NMR measurements, are similar to regions traditionally defined as disordered[2], we tested whether they are also identifiable with existing disorder predictors. To compare the disorder predictions with the two-dimensional ensembles in PODD, we introduced a cut-off value on the populations. We therefore defined as disordered those regions comprising at least L consecutive residues with a population of both α-helix and β-strand smaller than 0.5, and we calculated the balanced accuracy of the various predictors for different values of L (**Fig. 1d**). The resulting accuracies are not significantly different from those observed for a larger data set in which disorder was defined primarily from regions of missing electron density[24], suggesting that conventional binary definitions of order and disorder are contained within the continuum quantification provided by PODD.

Box 1 | Protein structural ensembles

In statistical mechanics, an ensemble can be defined as the set of all the states of a system together with their statistical weights (i.e., their populations). This type of description is often adequate to describe proteins in solution, both *in vitro* and *in vivo*, at least when they are not undergoing changes (for example, during chemical reactions). By adopting this view, the structural ensemble of a protein may be defined as the probability distribution of its possible conformations, each described, for instance, by its atomic coordinates. Other definitions are also possible, in which a conformation is defined through its native contacts or its secondary structure elements. In a structural ensemble at equilibrium, a state accessible to a protein is populated according to its Boltzmann weight ($\exp[-E/k_BT]/Z$), where E is the energy of the state, T the temperature, k_B is the Boltzmann constant, and Z is the partition function, which results from the balance between maximizing entropy and minimizing energy. Thus, in this framework, it is apparent that disordered regions are far from being random, as the statistical weights of the many conformations that they populate are typically very different depending on the energy of each conformation. Throughout this Commentary, we refer to 'protein dynamics' to indicate that proteins populate structural ensembles, as we are primarily referring to equilibrium properties of proteins ('equilibrium dynamics'), and we only touch in passing on non-equilibrium properties whose quantification requires a knowledge of the transitions rates between the populated states.



This figure provides a scheme of the different time scales (*x*-axis) and length scales (*y*-axis) probed by various methods that can be employed for the modeling of structural ensembles (top). Molecular dynamics (MD) methods are shown on a red background. X-ray crystallography and electron microscopy (EM) can also be used to investigate the different states populated by proteins[25]. Small-angle X-ray scattering (SAXS) and neutron scattering (SANS), as well as atomic force microscopy (AFM) and optical tweezers, can probe dynamics beyond the millisecond timescale and reveal the presence of different conformations of large complexes. We also show the different timescales that can be probed with various NMR techniques discussed in the main text (bottom). CS, chemical shifts; RDC, residual dipolar couplings; CPMG, Carr–Purcell–Meiboom–Gill.

Current challenges and opportunities

The growing arsenal of available NMR techniques is making it possible to study molecular systems of increasing complexity. Quantitative descriptions of protein equilibrium dynamics, such as the PODD annotations, can be used to readily infer the functional states of the proteins under scrutiny. For example, in the case of the cardiac isoform of troponin I (**Fig. 2a**), the secondary structure populations reveal how the presence of a binding partner shifts the equilibrium between the ordered and disordered states. This type of analysis can be used to identify and structurally characterize functionally relevant regions. Additionally, advances in in-cell NMR spectroscopy are making it increasingly possible to study protein structure and dynamics in bacteria as well as in mammalian cells[20,21]. The chemical-shift analysis employed in PODD can readily be applied in these studies, allowing fast structural investigation of proteins in their biological context. Furthermore, secondary structure populations determined this way can be compared with those observed in more controlled *in vitro* experiments to pinpoint the relevant states in cell. For example, the secondary-structure populations of α-synuclein in cell can be readily matched with those measured for the monomeric protein bound to sodium lauroyl sarcosinate (SLAS) micelles and in isolation *in vitro* (**Fig. 2b**). This comparison shows that the two-dimensional ensemble of α-synuclein in cell is essentially identical to that of the monomeric protein *in vitro*, as was also recently observed in mammalian cells with other NMR measurements[20]. Furthermore, it is becoming possible to use NMR spectroscopy to probe the dynamics of complex macromolecular systems, such as ribosome–nascent chain complexes[11]. A recent study of the cotranslational folding of an immunoglobulin-like domain has shown that the ribosome–nascent chain contains β-strands only marginally less stable than

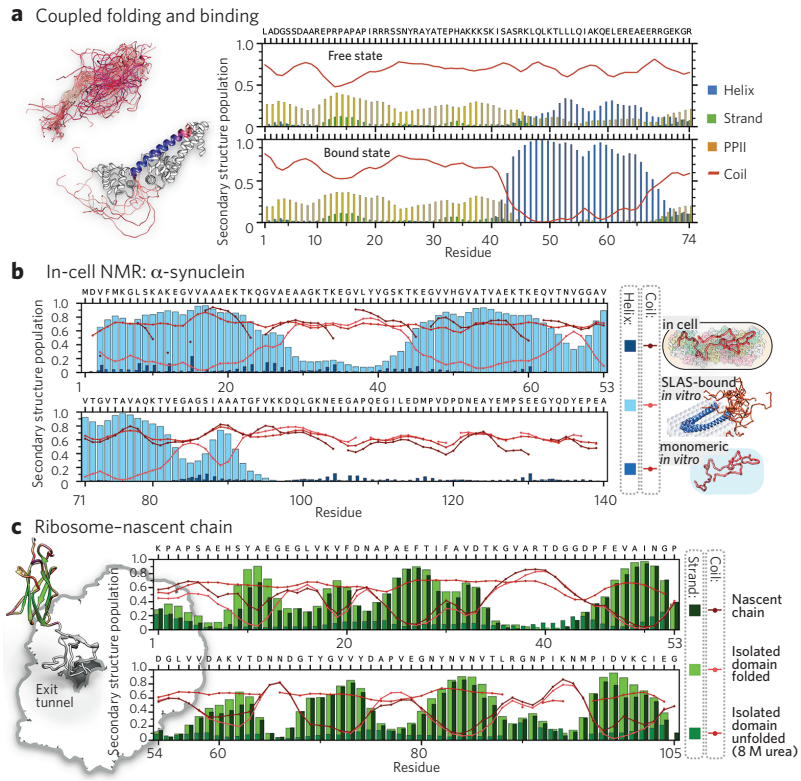


Figure 2 | Simultaneous quantification of protein order and disorder through two-dimensional ensembles. (**a**) N-terminal region of the cardiac isoform of troponin I ($cTnI_{[1-73]}$) in solution (top panel BMRB ID bmr25118) and bound to cardiac troponin C (cTnC, lower panel, bmr25119). (**b**) α-helix and random coil populations of α-synuclein from in-cell NMR experiments[21] (bmr19257) compared to those of the purified protein as a monomer in solution (bmr6968), and bound to SLAS micelles (bmr16302; see legend in the figure). Missing points in the ensembles correspond to residues without assigned chemical shifts. This analysis shows that α-synuclein populates in-cell states more similar to those of its monomeric disordered state than to the membrane-bound one, fully consistent with recent findings[20]. (**c**) β-strand and random coil populations of an immunoglobulin-like domain when part of a ribosome-nascent chain complex (bmr25748), compared to those of the isolated domain in its native and denatured states (bmr15814, see legend)[11].

those of the folded domain in isolation, indicating that this domain is essentially folded despite being tethered to the ribosome[11] (**Fig. 2c**).

Perspectives

In our opinion, it is time to take on the challenge of developing increasingly powerful quantitative structural methods and annotations for the effective representations of the dynamics of proteins in solution. The PODD database described above represents a step in this direction by providing a quantitative annotation that encompasses structure and equilibrium dynamics through the definition of secondary structure populations. We anticipate that in the near future it will be possible to further extend the amount of information conveyed by such annotations, as well as to increase their accuracy. A viable strategy may be to incorporate more sources of experimental data also capable of directly probing tertiary contacts, thus gradually converging toward methods of structural ensemble determination and integrative structural biology. A complementary strategy, which does not require additional experiments, is to integrate more *a priori* knowledge. This approach can be implemented, for instance, by exploiting the growing amount of available structural data or the increasingly accurate force fields, as is currently done by methods of structure prediction from NMR chemical shifts[16,18]. We thus suggest that the use of NMR spectroscopy, particularly in combination with other emerging experimental and computational approaches, will progressively enable large-scale quantitative structural and dynamic characterizations of proteins to be performed. The ability to simultaneously incorporate structure and dynamics in a unified framework will increase our understanding of the biological roles of order and disorder in proteins and will provide additional opportunities to identify key mechanisms of function, interaction, and allosteric regulation, as well as novel avenues for drug discovery.

[1]*Department of Chemistry, University of Cambridge, Cambridge , UK.* [2]*Department of Biomedical Sciences and CRIBI Biotechnology Center, University of Padova, Padova, Italy.* [3]*Department of Chemistry and Institute for Advanced Study, Technische Universität München, Garching, Germany.* [4]*MTA-DE Momentum Laboratory of Protein Dynamics, Department of Biochemistry and Molecular Biology, University of Debrecen, Hungary.* [5]*MTA-ELTE Lendület Bioinformatics Research Group, Department of Biochemistry, Eötvös Loránd University, Budapest, Hungary.* [6]*Department of Biomedical Engineering and Center for Biological Systems Engineering, Washington University in St. Louis, St. Louis, Missouri, USA.* [7]*MRC Laboratory of Molecular Biology, Cambridge, UK.* [8]*Architecture et Fonction des Macromolécules Biologiques (AFMB), Aix-Marseille Université, CNRS, Marseille, France.* [9]*VIB Center for Structural Biology, Vrije Universiteit Brussel, Brussels, Belgium, and Institute of Enzymology, Budapest, Hungary.* [10]*Center for Computational Biology and Bioinformatics, Department of Biochemistry & Molecular Biology, Indiana University Schools of Medicine & Informatics, Indianapolis, Indiana, USA.* [11]*Department of Molecular Biology and USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, Florida, USA.* **e-mail: mv245@cam.ac.uk*

References

1. Habchi, J., Tompa, P., Longhi, S. & Uversky, V.N. *Chem. Rev.* **114**, 6561–6588 (2014).
2. van der Lee, R. *et al. Chem. Rev.* **114**, 6589–6631 (2014).
3. Bhowmick, A. *et al. J. Am. Chem. Soc.* **138**, 9730–9742 (2016).
4. Dyson, H.J. & Wright, P.E. *Nat. Rev. Mol. Cell Biol.* **6**, 197–208 (2005).
5. Toth-Petroczy, A. *et al. Cell* **167**, 158–170.e12 (2016).
6. Fenwick, R.B., van den Bedem, H., Fraser, J.S. & Wright, P.E. *Proc. Natl. Acad. Sci. USA* **111**, E445–E454 (2014).
7. van den Bedem, H. & Fraser, J.S. *Nat. Methods* **12**, 307–318 (2015).
8. Ward, A.B., Sali, A. & Wilson, I.A. *Science* **339**, 913–915 (2013).
9. Lindorff-Larsen, K., Best, R.B., Depristo, M.A., Dobson, C.M. & Vendruscolo, M. *Nature* **433**, 128–132 (2005).
10. Baldwin, A.J. & Kay, L.E. *Nat. Chem. Biol.* **5**, 808–814 (2009).
11. Cabrita, L.D. *et al. Nat. Struct. Mol. Biol.* **23**, 278–285 (2016).
12. Jensen, M.R., Ruigrok, R.W. & Blackledge, M. *Curr. Opin. Struct. Biol.* **23**, 426–435 (2013).
13. Varadi, M. *et al. Nucleic Acids Res.* **42**, D326–D335 (2014).
14. Wüthrich, K. *Angew. Chem. Int. Ed. Engl.* **42**, 3340–3363 (2003).
15. Camilloni, C. *et al. Proc. Natl. Acad. Sci. USA* **111**, 10203–10208 (2014).
16. Cavalli, A., Salvatella, X., Dobson, C.M. & Vendruscolo, M. *Proc. Natl. Acad. Sci. USA* **104**, 9615–9620 (2007).
17. De Simone, A., Aprile, F.A., Dhulesia, A., Dobson, C.M. & Vendruscolo, M. *eLife* **4**, e02777 (2015).
18. Shen, Y. *et al. Proc. Natl. Acad. Sci. USA* **105**, 4685–4690 (2008).
19. Camilloni, C., De Simone, A., Vranken, W.F. & Vendruscolo, M. *Biochemistry* **51**, 2224–2231 (2012).
20. Theillet, F.-X. *et al. Nature* **530**, 45–50 (2016).
21. Waudby, C.A. *et al. PLoS One* **8**, e72286 (2013).
22. Bonomi, M., Camilloni, C., Cavalli, A. & Vendruscolo, M. *Sci. Adv.* **2**, e1501177 (2016).
23. Sormanni, P., Camilloni, C., Fariselli, P. & Vendruscolo, M. *J. Mol. Biol.* **427**, 982–996 (2015).
24. Walsh, I. *et al. Bioinformatics* **31**, 201–208 (2015).
25. Callaway, E. *Nature* **525**, 172–174 (2015).

Acknowledgments

This work was funded in part by COST ACTION bm1405 NGP-net.

Competing financial interests

The authors declare no competing financial interests.