Simultaneous quantification of protein order and disorder using NMR spectroscopy

Pietro Sormanni, Damiano Piovesan, Gabriella T Heller, Massimiliano Bonomi, Predrag Kukic, Carlo Camilloni, Monika Fuxreiter, Zsuzsanna Dosztanyi, Rohit Pappu, M Madan Babu, et al.

Simultaneous quantification of protein order and disorder using NMR spectroscopy

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 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], and Michele Vendruscolo[1,*]

Nuclear magnetic resonance spectroscopy is transforming our view of proteins by revealing how their structures and dynamics are closely intertwined to underlie their functions and interactions. Effective descriptions of protein dynamics are uncovering the presence and biological relevance of highly heterogeneous conformational states of proteins, which go beyond the traditional dichotomy between order and disorder by spanning the continuum between them.

The discovery of disordered proteins, which constitute about one third of the human proteome and are crucial for regulation and signalling[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 breakthrough has been followed by a polarisation of the terms 'order' and 'disorder', which, in hindsight, can be realized as largely prompted by a lack of experimental techniques capable of fully characterising protein dynamics. Protein disorder was thus at first defined as 'absence of structure', for example from missing electron densities in 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. We believe that this rather artificial polarisation will progressively disappear through the promotion and development of quantitative methods, such as nuclear magnetic resonance (NMR) spectroscopy, capable of simultaneously determining structure and dynamics (see **Box**).

A continuum between order and disorder. It is increasingly recognised that the native states of proteins range from being fully ordered to being almost completely disordered, with all the intermediate situations in between (**Figure 1**). In this context, it is becoming generally

accepted that the functional interpretation of structural results is often complicated by the fact that standard approaches have been specifically developed to define only the most representative 'static' structures within the ensembles populated in solution. Advances in kinetic protein crystallography[4,5] and integrative structural biology[6,7] (i.e. the combination of various methods of structural determination) are providing atomic-resolution descriptions of the states sampled on the picosecond to nanosecond timescales (see **Box**). Approaches of this type have already shown that many proteins for which a tightly-packed crystal structure determined at cryogenic temperature is available, are actually quite dynamic, in particular in regions important for function, interactions, and allosteric regulation[8]. There is, however, an even greater need to further develop techniques, such as NMR spectroscopy, capable of accurately describing the motions of larger amplitudes and longer timescales that are typical of disordered proteins (see **Box**). As progress will be made in this direction, the opposition between 'order' and 'intrinsic disorder' will gradually be replaced by organic 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 brought by NMR methods capable of determining quantitatively the structural fluctuations of proteins[9], which offer powerful means to achieve the simultaneous characterisation 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[10]. In more recent years, however, it has been increasingly realised that NMR measurements, since they report on average values over the structural fluctuations of proteins, 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**). In particular NMR spectroscopy is playing a crucial role in providing structural information about states that are intrinsically highly dynamical and can not be crystallised[11,12]. It is also becoming increasingly possible to use NMR spectroscopy to determine transition rates between different states, thus opening the way to the description of non-equilibrium dynamic processes[12] (see **Box**).

Challenges in the determination of protein structural ensembles. Notwithstanding this optimism about the potential of NMR spectroscopy in the 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 conformational states that yield the measured averages. Also, these measured averages provide sparse information - often coming from different types of experiments - concerning for example only certain bond angles and certain distances, which needs to be integrated together 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 (e.g. force fields) to model structural ensembles[13]. 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[14,15], to the large amplitude motions in the conformationally heterogeneous states populated by disordered proteins[11].

Towards a repository of protein structural ensembles. Quite generally, any method of ensemble modelling represents a compromise between: (1) the quality of the resulting structural ensemble, in particular in terms of 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 instead 'uncertainty ensembles'. To address this aspect, the Protein Ensemble Database (PED)[16] has been recently compiled. However, the relatively small number of entries (currently 24), reflects the fact that accurate structural ensemble calculations remain highly demanding both in terms of 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. The availability of increasingly accurate experimental and theoretical methods as well as the rapid growth of computing power will lead to the creation of a large repository of structural ensembles, capable of describing the properties of proteins in solution more comprehensively than static structures.

Two-dimensional ensembles in terms of secondary structure populations. 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 disordered states of proteins. In this context, the Protein Order and Disorder Database (PODD, www-mvsoftware.ch.cam.ac.uk/index.php/podd) contains the secondary structure populations of about 5,000 proteins, determined directly from NMR chemical shifts using the δ2D method[17]. The structural ensemble of the human prion protein (**Figure 1a**), and the corresponding secondary structure populations (**Figure 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 for this test case. Interestingly, a large fraction of residues catalogued in PODD is found in heterogeneous regions of proteins that significantly populate both α-helices and β-strands (**Figure 1c**). The main advantage in the use of secondary structure populations is that their determination is computationally rather inexpensive, and backbone chemical shifts are relatively 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[18].

The structural characterisation of proteins in PODD provides an illustration of the concept of a 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 continuum between them. To verify that the most dynamical regions present in PODD, where populations are derived from NMR measurements, are similar to regions traditionally defined disordered[2], we tested whether they are identifiable with existing disorder predictors. In order to compare the predictions with the two-dimensional ensembles in PODD we introduced a cut-off value by defining 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 (**Figure 1d**). The resulting accuracies are not significantly different from those observed on a larger dataset where disorder was defined primarily from regions of missing electron density[19], suggesting that conventional binary definitions of order and disorder are contained within the quantitative classification provided by PODD.

Current challenges and opportunities for NMR spectroscopy. The growing arsenal of available NMR techniques is making it possible to study molecular systems of great complexity. For instance, the PODD annotation can be used to readily infer functional states of the protein under scrutiny. For example, the presence of a binding partner shifts the equilibrium between the ordered and disordered states (**Figure 2a)**, which may be used to identify functionally relevant regions. In addition, great advances in *in cell* NMR spectroscopy make it increasingly possible to study protein structure and dynamics *in vivo* in bacteria as well as in mammalian cells[20,21]. The chemical shift analysis employed in PODD can readily be applied in these studies, thus allowing for fast structural investigation of proteins in their biological context. Furthermore *in cell* secondary structure populations can be compared with those observed in more controlled *in vitro* experiments in order to pinpoint the relevant states *in vivo*. For example, in **Figure 2b** we compare the secondary-structure populations of α-synuclein *in cell* with those measured for the monomeric protein bound to SLAS-micelles and in isolation *in vitro*. This comparison readily show that the two-dimensional ensemble of α-synuclein *in cell* is essentially identical to that of the monomeric protein in isolation, as also recently observed in mammalian cells with many measurements besides chemical shifts[21]. Furthermore, it is becoming possible to use NMR to probe the dynamics of complex macromolecular systems, such as ribosome-nascent chain complexes[11]. A recent study of the co-translational folding of an immunoglobulin-like domain has shown that the ribosome-nascent chain contains β-strands only marginally less stable than those of the folded domain in isolation, indicating that this domain is essentially folded on the ribosome[11] (**Figure 2c**).

Perspectives. In our opinion it is time to take on the challenge of introducing increasingly powerful quantitative structural methods and annotations that can lead to 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 the incorporation of more sources of experimental data, also capable of directly probing tertiary contacts, thus gradually converging towards methods of structural ensemble determination and integrative structural biology. A complementary strategy that does not require additional experiments is to integrate more *a priori* knowledge, for instance

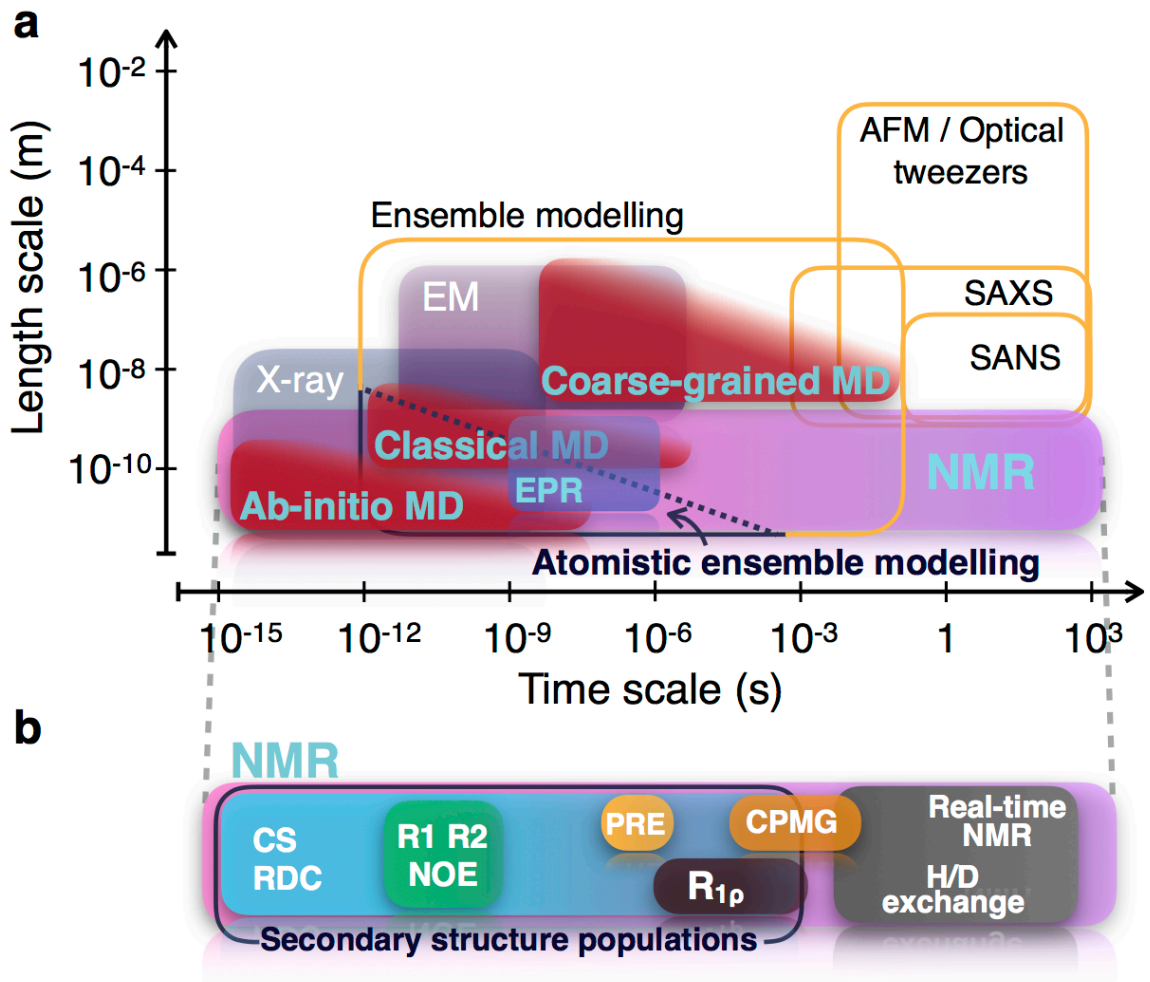
by exploiting the growing amount of structural data or the increasingly accurate force fields available, as currently done by methods of structure prediction from NMR chemical shifts[22,23]. We thus suggest that the use of NMR spectroscopy, in particular in combination with other emerging experimental and computational approaches, will progressively enable researchers to perform large-scale quantitative structural and dynamical characterisations of proteins. 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 regions for function, interactions, and allosteric regulation.

References

1. Habchi, J., Tompa, P., Longhi, S. & Uversky, V.N. Introducing protein intrinsic disorder. *Chem. Rev.* **114**, 6561-6588 (2014).
2. van der Lee, R. et al. Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **114**, 6589-6631 (2014).
3. Bhowmick, A. et al. Finding our way in the dark proteome. *J. Am. Chem. Soc.* **138**, 9730-9742 (2016).
4. Bourgeois, D. & Royant, A. Advances in kinetic protein crystallography. *Curr. Op. Struct. Biol.* **15**, 538-547 (2005).
5. Hajdu, J. et al. Analyzing protein functions in four dimensions. *Nat. Struct. Mol. Biol.* **7**, 1006-1012 (2000).
6. Fenwick, R.B., van den Bedem, H., Fraser, J.S. & Wright, P.E. Integrated description of protein dynamics from room-temperature X-ray crystallography and NMR. *Proc. Natl. Acad. Sci. USA* **111**, E445-E454 (2014).
7. Ward, A.B., Sali, A. & Wilson, I.A. Integrative structural biology. *Science* **339**, 913-915 (2013).
8. van den Bedem, H. & Fraser, J.S. Integrative, dynamic structural biology at atomic resolution - It's about time. *Nat. Methods* **12**, 307-318 (2015).
9. Lindorff-Larsen, K., Best, R.B., DePristo, M.A., Dobson, C.M. & Vendruscolo, M. Simultaneous determination of protein structure and dynamics. *Nature* **433**, 128-132 (2005).
10. Wüthrich, K. NMR studies of structure and function of biological macromolecules (Nobel lecture). *Angew. Chem. Int. Ed.* **42**, 3340-3363 (2003).
11. Cabrita, L.D. et al. A structural ensemble of a ribosome-nascent chain complex during cotranslational protein folding. *Nat. Struct. Mol. Biol.* **23**, 278-285 (2016).
12. Baldwin, A.J. & Kay, L.E. NMR spectroscopy brings invisible protein states into focus. *Nat. Chem. Biol.* **5**, 808-814 (2009).
13. Bonomi, M., Camilloni, C., Cavalli, A. & Vendruscolo, M. Metainference: A Bayesian inference method for heterogeneous systems. *Sci. Adv.* **2**, e1501177 (2016).
14. De Simone, A., Aprile, F.A., Dhulesia, A., Dobson, C.M. & Vendruscolo, M. Structure of a low-population intermediate state in the release of an enzyme product. *eLife* **4**, e02777 (2015).

15. Camilloni, C. et al. Cyclophilin a catalyzes proline isomerization by an electrostatic handle mechanism. *Proc. Natl. Acad. Sci. USA* **111**, 10203-10208 (2014).
16. Varadi, M. et al. pE-DB: A database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucl. Acids Res.* **42**, D326-D335 (2014).
17. Camilloni, C., De Simone, A., Vranken, W.F. & Vendruscolo, M. Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. *Biochemistry* **51**, 2224-2231 (2012).
18. Sormanni, P., Camilloni, C., Fariselli, P. & Vendruscolo, M. The s2d method: Simultaneous sequence-based prediction of the statistical populations of ordered and disordered regions in proteins. *J. Mol. Biol.* **427**, 982-996 (2015).
19. Walsh, I. et al. Comprehensive large-scale assessment of intrinsic protein disorder. *Bioinformatics* **31**, 201-208 (2015).
20. Waudby, C.A. et al. In-cell NMR characterization of the secondary structure populations of a disordered conformation of α-synuclein within E. coli cells. *PloS One* **8**, e72286 (2013).
21. Theillet, F.-X. et al. Structural disorder of monomeric α-synuclein persists in mammalian cells. *Nature* (2016).
22. Cavalli, A., Salvatella, X., Dobson, C.M. & Vendruscolo, M. Protein structure determination from NMR chemical shifts. *Proc. Natl. Acad. Sci. USA* **104**, 9615-9620 (2007).
23. Shen, Y. et al. Consistent blind protein structure generation from NMR chemical shift data. *Proc. Natl. Acad. Sci. USA* **105**, 4685-4690 (2008).
24. Lang, P.T., Holton, J.M., Fraser, J.S. & Alber, T. Protein structural ensembles are revealed by redefining X-ray electron density noise. *Proc. Natl. Acad. Sci. USA* **111**, 237-242 (2014).
25. Cossio, P. & Hummer, G. Bayesian analysis of individual electron microscopy images: Towards structures of dynamic and heterogeneous biomolecular assemblies. *J. Struct. Biol.* **184**, 427-437 (2013).

Box: Protein structural ensembles



In statistical mechanics, an ensemble is defined as the set of all the states of a system together with their statistical weights. 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 rapid changes (e.g. during chemical reactions). By adopting this view, the structural ensemble of a protein may be defined as the probability distribution of its members, each described for instance by its atomic coordinates relative to a fixed reference frame. Other definitions are also possible, where a structure is defined through its native contacts or its secondary structure elements. In a structural ensemble, disordered proteins or disordered regions are far from being random - they do populate a vast number of different states, but the statistical weights (i.e. the populations) of such states are also typically very different. In this commentary, we speak about 'protein dynamics' to indicate that proteins populate structural ensembles. We note, however, that we are primarily referring to equilibrium properties of proteins, and we only

touch in passing on non-equilibrium properties that depend on the transitions rates between the populated states.

The figure provides a schematic illustration of the different length scales (y-axis) and time scales (x-axis) that can be probed with various methods of studying protein structure and dynamics. Methods that do not yield atomistic resolution are framed in orange, computational methods are shown on a red background. X-ray crystallography and electron microscopy (EM) are shown in white as they are traditionally employed to obtain static structural information, even if recent applications demonstrated that they can be used to investigate the different states populated by proteins[24,25]. NMR spectroscopy can most effectively shed light on the dynamics of small to medium-sized proteins on a wide range of timescales. Chemical shifts (CS) and residual dipolar couplings (RDC), which span the range from femtoseconds to milliseconds, can probe biochemical processes ranging from bond vibrations and electron transfer, to protein folding, ligand binding, allostery, and catalysis. Other NMR measurements, such as those exploiting nuclear Overhauser effects (NOEs), and R1/R2 relaxation rates, which probe the picosecond to nanosecond timescales, are informative of hydrogen-bond formation, hydrogen transfer, side-chain rotation and rotational diffusion. Paramagnetic relaxation enhancement (PRE) experiments probe the dynamics in the microsecond time scale, typical of secondary structure formation and fast folding, unfolding and ligand binding processes. Electronic paramagnetic resonance (EPR) spectroscopy - a technique whose basic concepts are analogous to those of NMR spectroscopy - can also be applied to study timescales between the nanosecond and the microsecond. Extending beyond the microsecond into the millisecond time scale, NMR R1ρ rotating-frame relaxation and Car-Purcell-Meiboom-Gill (CPMG) data provide information into slow-folding/unfolding proteins and binding processes. On the longest time scale reachable by NMR techniques, real-time NMR and hydrogen/deuterium (H/D) exchange data probe dynamics up to the second timescale, typical of transport and protein translation.

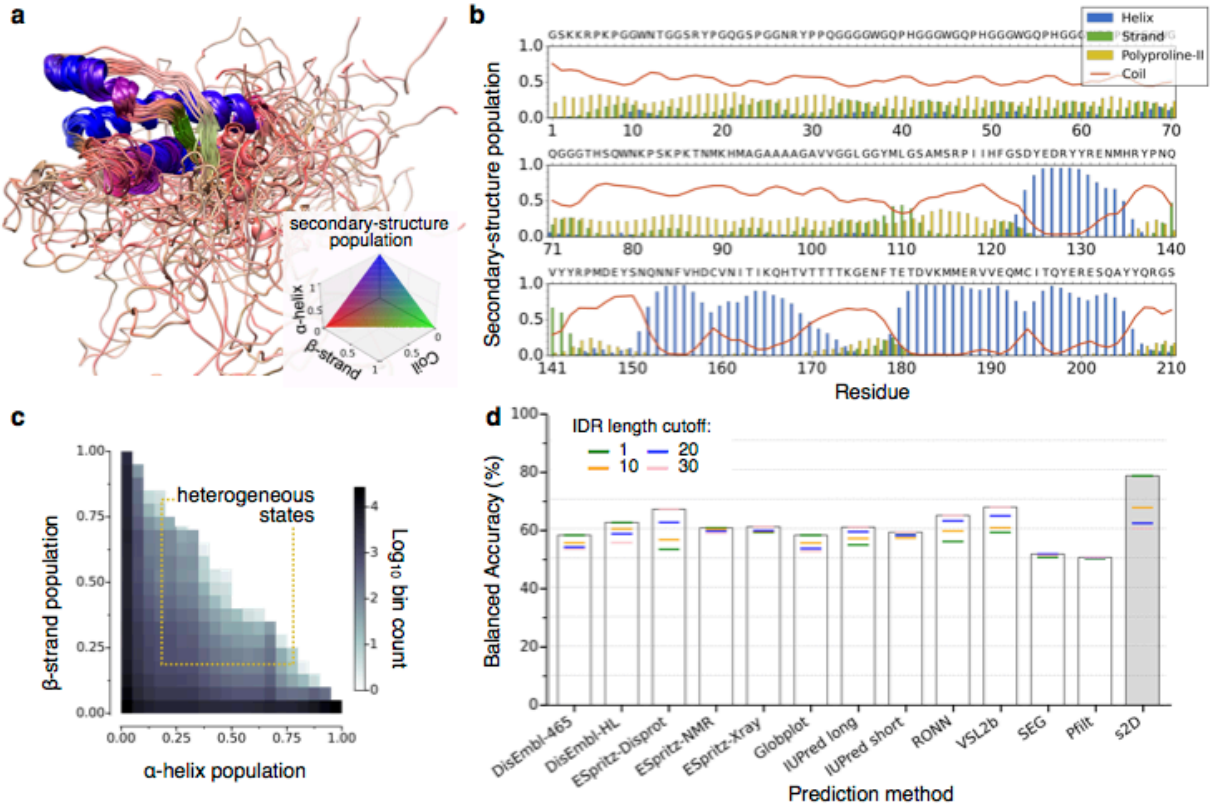


Figure 1. Structural ensemble of the human prion protein (**a**), and corresponding secondary structure populations (**b**) as calculated from NMR chemical shifts (BMRB ID 4402) using the metainference method[13]. (**c**) Scatter plot of the α-helix and β-strand populations for all residues in the PODD dataset. The dashed rectangle highlights residues in heterogeneous regions, which significantly populate more than one type of secondary structure element. (**d**) Bar plot of the balanced accuracy of sequence-based methods of predicting disorder (x-axis) on a subset of the PODD dataset 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 ordered otherwise. The column for the s2D method[18] is in grey as some sequences in the dataset are part of its training set.

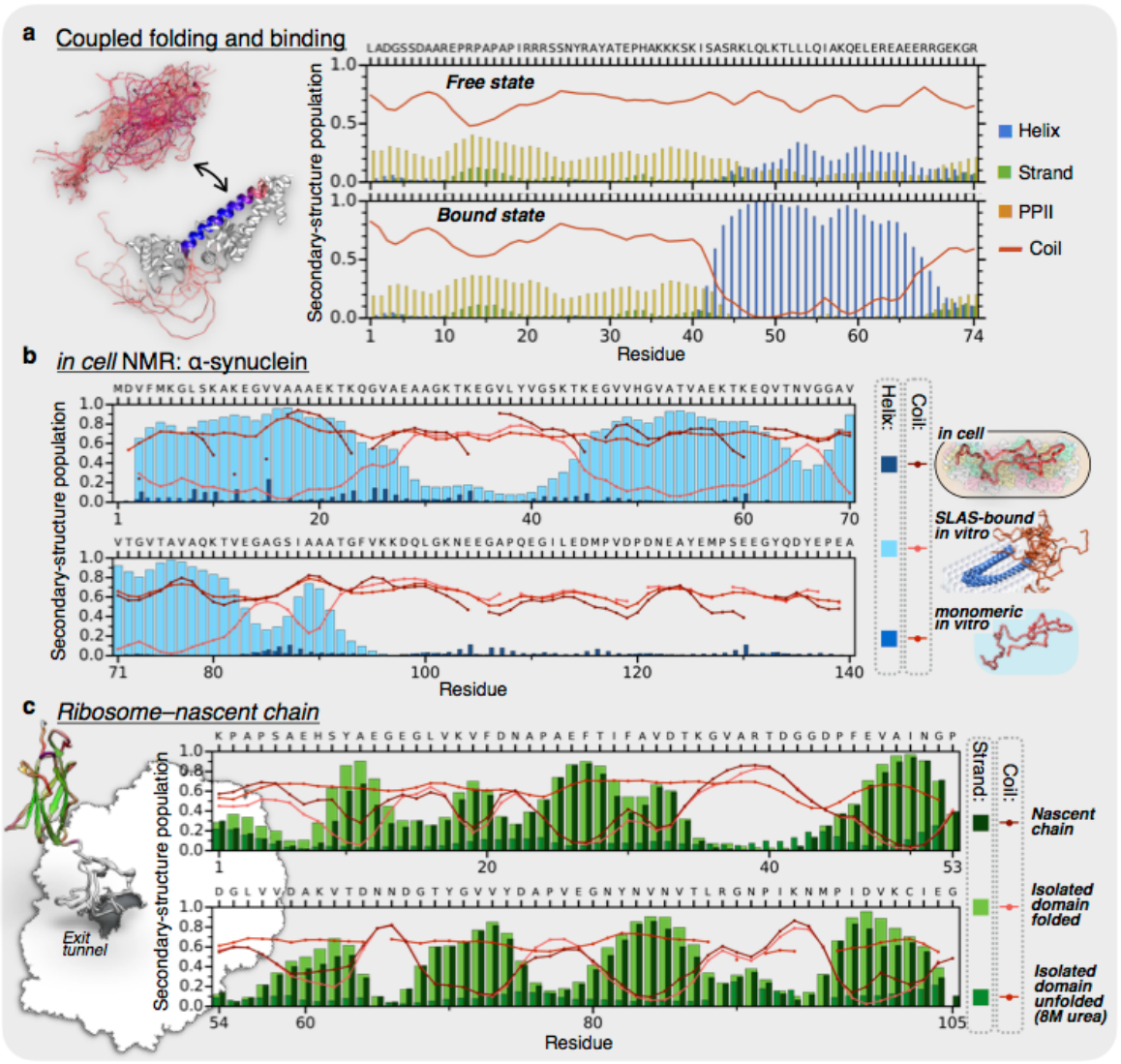


Figure 2. Example applications of two-dimensional ensembles. (**a**) N-terminal region of the cardiac isoform of troponin I (cTnI$_{[1\text{-}73]}$) in solution (top panel BMRB ID bmr25118) and bound to cardiac troponin C (cTnC - lower panel, bmr25119). (**b**) α-helix and random coil populations of α-synuclein from an *in cell* NMR experiment[20] (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). This analysis shows that α-synuclein *in cell* populates states more similar to those of its monomeric disordered state than to the membrane-bound one, fully consistent with recent findings[21]; missing points in the ensembles correspond to residues without assigned chemical shifts. **(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].

Author affiliations

[1]Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK. [2]Department of Biomedical Sciences and CRIBI Biotechnology Center, University of Padova, 35121 Padova, Italy. [3]Department of Chemistry and Institute for Advanced Study, Technische Universität München, Lichtenbergstrasse 4, D-85747 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, MO 63130. [7]MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, United Kingdom. [8]Aix-Marseille Univ, CNRS, Architecture et Fonction des Macromolécules Biologiques (AFMB), [9]VIB Department of Structural Biology, Vrije Universiteit Brussel; Brussels, Belgium; Institute of enzymology; Budapest, Hungary; [10]Center for Computational Biology and Bioinformatics, Department of Biochemistry & Molecular Biology, Indiana University Schools of Medicine & Informatics, Indianapolis, IN, USA; [11]Department of Molecular Biology and USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, FL, USA

Acknowledgments

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