Predicting the Predictive Power of IDP Ensembles

Peter Tompa[1,2,*] and Mihaly Varadi[1]
[1]VIB Department of Structural Biology, Vrije Universiteit Brussel, 1050 Brussels, Belgium
[2]Institute of Enzymology, Research Centre for Natural Sciences of the Hungarian Academy of Sciences, 1113 Budapest, Hungary
*Correspondence: ptompa@vub.ac.be
http://dx.doi.org/10.1016/j.str.2014.01.003

The function of intrinsically disordered proteins may be interpreted in terms of their structural ensembles. The article by Schwalbe and colleagues in this issue of *Structure* combines NMR and SAXS constraints to generate structural ensembles that unveil important functional and pathological features.

The major aspiration of the structure-function paradigm is to interpret protein function at atomic detail based on three-dimensional protein structures. A recent shift in this paradigm has been provoked by the recognition that intrinsically disordered proteins (IDPs), or regions (IDRs), exist and function without a well-defined structure. The phenomenon of structural disorder is prevalent in proteins of signaling and regulatory functions and is also frequently involved in diseases, such as cancer or neurodegenerative disorders (Tompa, 2012).

Despite the recognition of its importance, for almost a decade, the field of protein disorder could not progress beyond the general statement that function is compatible with the lack of a well-defined structure. Our recent aim has been to describe structures as a collection of a large number of conformations (an ensemble) and interpret function in terms of its characteristic features, as formulated in the call for "unstructural" biology (Tompa, 2011).

The description of IDPs by ensembles is nontrivial, because the number of degrees of conformational freedom is much greater than the number of experimental observables that can be determined. The approaches recently developed to address this inherently ill-posed problem are primarily based on nuclear magnetic resonance (NMR) data combined with additional restraints obtained from small-angle X-ray scattering (SAXS) (Fisher and Stultz, 2011) either by running constrained molecular dynamics simulations where the conformational sampling is biased by experimental restraints (Allison et al., 2009) or by the selection of a limited number of conformations from a very large random pool that can describe the experimental data (Jensen et al., 2010; Ozenne et al., 2012a) (Figure 1). Because of the underdetermined nature of the problem, several ensembles fit the data equally well. In an attempt to counter the degeneracy of ensembles, every method tries to integrate data sensitive to (1) short range structural order, such as chemical shifts (CSs), residual dipolar couplings (RDCs), J-couplings, hydrogen-exchange protection factors, relaxation rates, and solvent-accessibility and (2) long range structural order, such as paramagnetic relaxation enhancements (PREs), nuclear Overhauser effects (NOEs), hydrodynamic parameters, and SAXS topological restraints. The ensembles with the best fit to the data are made public via deposition into the Protein Ensemble Database (Varadi et al., 2014).

However, ensembles with an equally good fit cannot yet be distinguished; it has not been assessed if they are representative of the entire conformational space and/or if they can describe important functional and/or pathological features (Fisher and Stultz, 2011). These critical issues are now addressed by solving and analyzing the ensemble of α-synuclein and tau protein in an article in this issue of *Structure* (Schwalbe et al., 2014). These proteins are involved in neurodegenerative disorders where they convert from the soluble, disordered physiological state to an insoluble, pathological amyloid form dominated by β structures.

α-synuclein and tau protein have been selected to serve as test cases for solving and benchmarking IDP ensembles, because: (1) the size of tau protein is 441 residues, posing methodological challenges associated with long IDPs; (2) their global structural features are already known: an extended structure combined with long-range interactions and propensity to sample compact states; (3) localized functional interaction regions (tubulin-binding regions of tau protein) are known; and (4) their involvement in neurodegenerative diseases, with known regions initiating coil-to-beta transition. Dissecting all these functionalities is the major challenge in interpreting the ensembles. By combining a large number of NMR- and SAXS-derived constraints, Schwalbe et al. (2014) calculated representative ensembles of 200 (α-synuclein) and 400 (tau protein) structures and devised quantitative measures of their structural predictive power (Figure 1). It was shown that the ensembles can predict independent experimental observables and suggest local conformational features potentially involved in function and diseases.

The underlying experimental data included 5 NMR CS values, 3 RDC values, 12 PRE measurements, and SAXS scattering data. Flexible meccano was used to generate a large number of statistical coils (Ozenne et al., 2012a), followed by the genetic algorithm of ASTEROIDS (Jensen et al., 2010) to select ensembles compatible with experimental data. For the first time, the ensemble descriptions are crossvalidated with independent experimental data, which provides a quantitative measure of their predictive capacity (Figure 1). The ensembles selected show signs of predictive power in several aspects. (1) Non-random behavior: Different combinations of data (CSs and and/or RDCs) were removed from the analysis in all five cases (full-length proteins and tau segments) and were found to be predicted more accurately by the ensemble than by statistical-coil descriptions. The improvement increased in regions where local sampling deviates

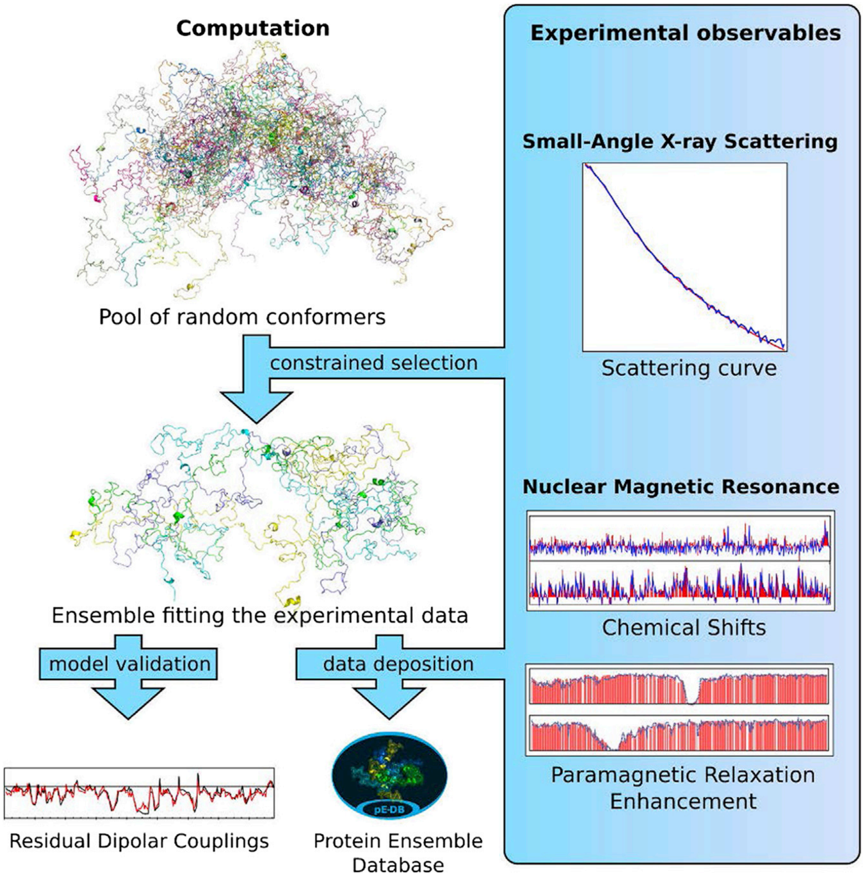
Computation
Pool of random conformers
constrained selection
Ensemble fitting the experimental data
model validation
data deposition
Residual Dipolar Couplings
pE-DB
Protein Ensemble Database
Experimental observables
Small-Angle X-ray Scattering
Scattering curve
Nuclear Magnetic Resonance
Chemical Shifts
Paramagnetic Relaxation Enhancement

Figure 1. Workflow of Solving Ensembles

Descriptive structural ensembles of IDPs/IDRs can be solved by a clever interplay of dedicated experimental observation (NMR, SAXS, and possibly other techniques) and computational tools. As suggested and demonstrated in the Schwalbe et al. (2014) paper, by combining with evaluation supported by data deposition, ensembles can predict critical structural and functional features of disordered proteins.

from statistical coil. (2) Correlation with function: the combination of CSs and RDCs has been shown to be able to distinguish between local populations of secondary structure (Ozenne et al., 2012b), which might be very important because IDPs/IDRs often locally presample their bound conformation in solution (Tompa, 2012). In tau protein, the four Gly-rich sequences in repeat domains involved in microtubule-binding, as well as four type I β turns, have significantly increased α population. Flanking these regions are polyproline II (βP) stretches, which may be important in exposing these regions for interaction. (3) Correlation with disease: both proteins show an elevated population of βP conformation, which might support the hypothesis that the βP region of conformational space represents a precursor for aggregation and formation of stable β sheets (Blanch et al., 2000). This sampling is localized in the vicinity of the aggregation nucleation sites of the proteins, such as aggregation nucleation hexapeptides in tau protein and the NAC region of α-synuclein.

All these correlations support the point that the ensemble description of IDPs/IDRs might have the power of elucidating the functional features of disordered proteins. Can we claim to have come close with IDPs to the success of structural biology of ordered proteins? Definitely not; there is still a long way to go for the maturation of unstructural biology (Tompa, 2011). Among other things, we need to (1) exploit other types of experimental data such as fluorescence resonance energy transfer, electron paramagnetic resonance, or mass spectroscopy (Schwalbe et al., 2014); (2) try to distinguish between equivalent ensembles by improving calculation skills; (3) combine ensemble structural data with diverse functional data (e.g., evolutionary information and mutagenesis); (4) make ensembles generally available for the community for critical evaluation (Varadi et al., 2014); and (5) include the fourth dimension of structure—dynamics—in ensemble descriptions. Thanks to ground-breaking work (Schwalbe et al., 2014), progress in all of these areas is anticipated to bear fruit in the near future.

REFERENCES

Allison, J.R., Varnai, P., Dobson, C.M., and Vendruscolo, M. (2009). J. Am. Chem. Soc. *131*, 18314–18326.

Blanch, E.W., Morozova-Roche, L.A., Cochran, D.A., Doig, A.J., Hecht, L., and Barron, L.D. (2000). J. Mol. Biol. *301*, 553–563.

Fisher, C.K., and Stultz, C.M. (2011). Curr. Opin. Struct. Biol. *21*, 426–431.

Jensen, M.R., Salmon, L., Nodet, G., and Blackledge, M. (2010). J. Am. Chem. Soc. *132*, 1270–1272.

Ozenne, V., Bauer, F., Salmon, L., Huang, J.R., Jensen, M.R., Segard, S., Bernadó, P., Charavay, C., and Blackledge, M. (2012a). Bioinformatics *28*, 1463–1470.

Ozenne, V., Schneider, R., Yao, M., Huang, J.R., Salmon, L., Zweckstetter, M., Jensen, M.R., and Blackledge, M. (2012b). J. Am. Chem. Soc. *134*, 15138–15148.

Schwalbe, M., Ozenne, V., Bibow, S., Jaremko, M., Jaremko, L., Gajda, M., Jensen, M.R., Biernat, J., Becker, S., Mandelkow, E., et al. (2014). Structure *22*, this issue, 238–249.

Tompa, P. (2011). Curr. Opin. Struct. Biol. *21*, 419–425.

Tompa, P. (2012). Trends Biochem. Sci. *37*, 509–516.

Varadi, M., Kosol, S., Lebrun, P., Valentini, E., Blackledge, M., Dunker, A.K., Felli, I.C., Forman-Kay, J.D., Kriwacki, R.W., Pierattelli, R., et al. (2014). Nucleic Acids Res. *42*, D326–D335.