Folding proteins one loop at a time.

Daan Frenkel, Cambridge

My least favourite opening sentence of a scientific article is: "Recently there has been much interest in…". Unless the sentence goes on to explain why there is so much interest in the topic, it is unlikely to excite the curiosity of the reader.

Let me, nevertheless, start this Commentary with the words: "Over the past few decades there has been much interest in protein folding…". But, obviously, this statement has to be followed by the question: "Why?".

Interestingly, there are at least two different yet valid answers to this question. The simplest is that if we know how proteins fold, then we can predict the three-dimensional structure of a functional protein on the basis of its amino-acid residue sequence alone. At present, successful predictions of the structure of all but the smallest proteins are not – or at the very least, not exclusively based – on molecular modelling. The `data-mining' techniques that are being used instead are sophisticated and fascinating (see, e.g. [1]), but they rarely deal with the question *how* proteins fold: the focus is on the final product.

The second motivation for studying protein folding is to understand how a protein can reach its target state from an initially disordered state [2-5]. Of course, many proteins do not normally fold on their own – they may be protected and helped by chaperones and similar molecules, but interestingly, many proteins can fold spontaneously if the need arises. Hence, the folding ability of proteins has been programmed into their sequence. This is a very interesting phenomenon, because it means that proteins are their own assembly manual. Clearly, now that we are moving towards complex, man-made macro-molecular structures (e.g. based on DNA origami), we would like to learn from the instruction set that proteins carry. Knowledge of the pathway(s) along which proteins fold is also interesting when we wish to understand under what circumstances proteins folding goes wrong, causing diseases such as Parkinson's, Alzheimer's or type-II diabetes. Understanding how to control protein folding is likely to be an essential ingredient in future intervention strategies.

The article of Jacobs and Shakhnovich [6] in this issue provides important new insights in possible pathways that proteins traverse on their way from the initial disordered state to the folded state. In particular, the work presents a computationally tractable model that predicts all intermediate structures that are likely to form as folding proceeds: i.e the focus is on the whole path, not just on the transition state. The strength of the model is that it allows for the fact that several non-adjacent ordered domains in a partially folded protein may be present simultaneously. Existing simple models lacked the combination of simplicity and realism to account for the existence of intermediate structures with an arbitrary number of ordered subdomains (for an early model study, see [7]) – and fully atomistic models tend to be depressingly expensive.

The model proposed by Jacobs and Shakhnovich describes a protein as a graph with a single backbone. The model only considers those non-bonded interactions between amino-acid residues that are present in the native state (hence, knowledge of the structure of the native state is essential). In the model of Jacobs and Shakhnovich, folding proceeds from the disordered state where none of the native-state interactions have formed towards the folded (native) state by forming one ordered sub-domain after the other. However, the order in which these domains form is not fixed – hence the model can account for many parallel folding pathways (see Figure).

The crucial observation is that the key activated steps in protein folding are the moments when a new domain starts to form: the `gluing together' of two parts of the protein that were initially able to move with respect to each other. The formation of a new ordered domain changes the topology of the protein. Clearly this step costs entropy, and it is this entropic cost that is responsible for the free energy barrier that separates a partially folded protein with **n** and **n+1** ordered domains. Crucially, the model makes it possible to tell which residues are most likely to be involved in the contacts that form when a protein changes its topology.

Key results of the study, which contains comparison with proteins that have been well studied in very long atomistic simulations [8], are that there is a limited number of dominant pathways that proceed via relatively long lived intermediate structures with differing numbers of ordered domains. Moreover, the simulations show which domains are involved in the folding process and which residues determine the barriers that separate different topological states of the folding proteins.

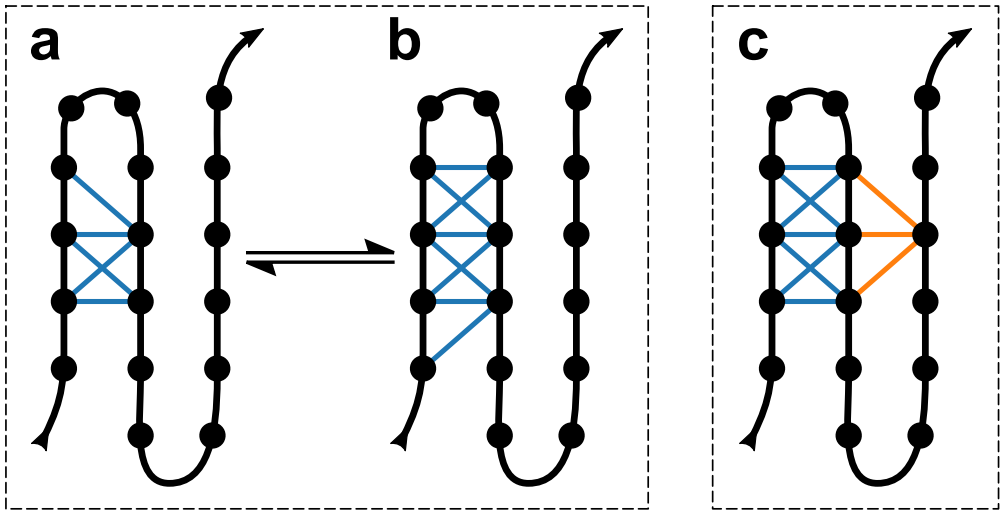
The computational approach used by Jacobs and Shakhnovich is based on an extension of the approach used in ref. [9]. The power of the method should extend well beyond protein folding. An example where the method could be applied is in predicting – and eventually improving - the `folding pathway' of DNA origami structures. Similarly, the method should be well suited to predict RNA folding, including the folding of structures that contain pseudo-knots.

Finally, it is interesting to speculate how this approach could be used to understand, and eventually control, protein misfolding. In the spirit of the work of Jacobs and Shakhnovich, misfolding pathways may be very different from folding pathways, as the contacts in a fully formed misfolded structure are different from the native contacts. However, once these contacts are known from experiment, it becomes possible to predict the misfolding pathway, to identify the long-lived intermediates and, crucially, to predict which residues are involved in the rate limiting topology-changing steps. One can well imagine that this knowledge could suggest targets for drugs that could inhibit misfolding.

[1] Marks, D.S., Colwell., J.L., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., Sander, C. (2011), Protein 3D Structure Computed from Evolutionary Sequence Variation, PLoS ONE 6, e28766.
[2] Sali, A., E. I. Shakhnovich, and M. Karplus, 1994. How does a protein fold? Nature 369:19.
[3] Onuchic, J. N., and P. G. Wolynes, 2004. Theory of protein folding. Curr. Opin. Struct. Biol. 14:70–75.

[4] Daggett, V., and A. Fersht, 2003. The present view of the mechanism of protein folding. Nat. Rev. Mol. Cell Biol. 4:497–502.
[5] Jackson, S. E., and A. R. Fersht, 1991. Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition. Biochemistry 30:10428–10435.
[6] Jacobs, W. M., and Shakhnovich, E. I., 2016, Structure-based prediction of protein-folding transition paths, Biophys. J. 306420R.
[7] Munõz, V., and W. A. Eaton, 1999. A simple model for calculating the kinetics of protein folding from three-dimensional structures. Proc. Natl. Acad. Sci. U.S.A. 96:11311–11316.
[8] Piana, S., K. Lindorff-Larsen, and D. E. Shaw, 2013. Atomic-level description of ubiquitin folding. Proc. Natl. Acad. Sci. U.S.A. 110:5915–5920.
[9] Jacobs, W. M., A. Reinhardt, and D. Frenkel, 2015. Communication: Theoretical prediction of free-energy landscapes for complex self-assembly. J. Chem. Phys. 142:021101.

Figure:



CAPTION:
Example of the schematic representation of a protein during foding. The black dots denotes bonds along the protein backbone. The colored lies denote native contacts that have formed during folding. Note that the topology of the protein if panels **a** and **b** is the same: the difference is only in the number of connected residues in the domain. In contrast, panel **c** shows a different topology where now two ordered domains are present (and an extra loop has formed). Note however that the number of bonded residues in situation **c** needs not be larger than in **a** or **b** (in fact, for the example shown, the number of bonded residues in **c** happens to be the same as in **b**).