Proteins are synthesized in the cell as sequential strings of amino acids. However, ribosome-emerging polypeptide chains are not active and they have to fold into a usually unique three-dimensional (3D) structure to become functional. Many proteins have been shown to attain spontaneously in vitro their native state from an initially complex ensemble of unfolded conformations in the absence of the cellular protein quality control. This implies that the information determining how proteins gain their active 3D structures should be imprinted somehow in their primary sequences. Nevertheless, the specific connection between protein sequence and function has remained lost for almost six decades in the black-box of protein folding. Understanding how linear amino acids chains are pre-programmed to fold into their specific functional architectures remains as one of the most challenging and important tasks in molecular biology. Now, in a recent theoretical work Mittal et al. put forward a simple principle to explain backbone organization in protein folding (1). Essentially, the authors propose the absence of preferential long-, medium- or short-range interactions between amino acids during the folding of globular proteins. It must be mentioned that preferential interactions between amino acids were the basis for introducing knowledge-based potentials, which in turn provided the underpinning for present day 3D protein structure prediction by modeling and simulation (2-4 and references therein).

The process of folding would be determined instead by the stoichiometric distribution of the twenty proteinogenic amino acids in the primary sequences of proteins in a similar way that the stoichiometry of the four nucleotides, or Chargaff’s rules, drive the spatial organization of the DNA double helix. The authors delineate these conclusions upon analyzing the distances between Cα atoms in 3718 proteins in the PDB concluding that the most important factors for the folding of a protein are exclusion by water, the shape of the individual amino acids (not their physicochemical properties) and the way these shapes are distributed along the sequence. There is no doubt that this is a revolutionary idea that if comes to be true would change our present view of protein folding and have many important implications in related areas like structural prediction, protein design or protein engineering.

Unfortunately, the theoretical framework proposed by Mittal et al. at least in its present formulation, clashes with different experimental evidences, like the folding behaviour of retro-proteins. Both the classical view and the new proposal agree that a polypeptide with exactly the same stoichiometry as a naturally occurring protein, but with a randomized distribution of residues or, in the authors’ formulation, shapes would not get the same native conformation and very likely would not fold at all. But what would happen if we read the primary sequence backwards generating thus a retro-protein? Because native proteins are usually not palindromic in sequence, this exercise would result in a new polypeptide

**Salvador Ventura**

Institut de Biotecnologia i de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain.

Corresponding Author: Salvador Ventura
Phone: 34-93-5868956
Fax: 34-93-5811264
E-mail: salvador.ventura@uab.es
that does not align with its parent sequence and therefore, according to the prevalent view of folding, it will not fold into the same 3D structure, in case it can fold. Instead, if one visualizes protein folding as proposed by Mittal et al. because they display identical stoichiometry and the same sequential distribution of amino acids shapes, their packing can potentially attain the same minimal surface-to-volume ratio and there is not obvious reason to doubt that the two protein will attain the same 3D structure. The reality is that, as a general rule, retro-proteins do not fold into the compact, stable and soluble conformations characteristic of natural globular proteins (5).

An important concept in the article by Mittal et al. is what the authors define as the “margin of life” or the specific distribution of amino acids compatible with folded proteins. The proposed stoichiometry displays low standard deviations, implying that significant divergences from this distribution would have deleterious structural effects. These so said Chargaff’s rules of proteins would impose thus strong compositional constraints for the design of new non-natural proteins. However, several protein design exercises have succeeded in producing new proteins using reduced amino acid alphabets. Although the sequences of these proteins deviate significantly in composition from the one proposed by Mittal et al. they encode for stable, and topologically complex native conformations which are able to fold in a biologically relevant time frame (6). In addition, as shown for the paradigmatic case of bovine pancreatic trypsin inhibitor, extensively simplified proteins in which over one-third of the residues are alanines retain the overall backbone and side-chain configurations of the natural inhibitor (7). Together, these observations support a view in which most protein structure determinants occur at a few sites and involve the establishment of selected side chain-side chain interactions, questioning thus the role of strict stoichiometric laws in protein folding.

The discrepancies between Mittal et al. calculations and the above mentioned experimental evidences likely arise from the fact that the authors measure distances between Cα atoms in already folded globular proteins. Unfortunately, this says little about the process of folding, since the fact that Cα atoms are separated in the native structure by a short distance does not necessarily implies that the corresponding residues would be interacting in the native conformation, nor that they established contacts during the process of protein folding, since they could be brought in close vicinity simply by the interactions established by neighbouring residues. This brings us to the concept of folding nucleus, defined as key residues in the sequence, which contacts can lead the folding of a polypeptide chain towards the acquisition of its unique functional structure. The number of residues involved in such interactions is usually small. In this way, for acylphosphatase it appears that an extensive long range native-like contact network established by only three residues during the folding reaction is sufficient to determine the overall conformation of the protein (8). Importantly, in the native structure these three key residues are not in contact. Therefore, approximations like the one of Mittal et al. will underscore their crucial role in folding, while providing more relevance to residues that are in the vicinity simply because the interaction between residues in the folding nucleus make them come together, not being necessary that they interact directly with each other during or after folding.

Overall, although experimental data do not provide strong evidences for stoichiometry being a major factor controlling the folding reactions of natural proteins, we should not dismiss the relevance of the work of Mittal et al. It is clear that the field needs new ways of addressing the folding problem that by generating debate on the mechanisms underlying this process might allow us to advance towards the formulation of general and simple rules to understand how proteins fold.

References