The article “A Stoichiometry Driven Universal Spatial Organization of Backbones of Folded Proteins: Are there Chargaff’s Rules for Protein Folding?” by Mittal et al. (1) seeks to relate the ability of linear amino acid sequences to form folded protein structures and the frequency of occurrence of those amino acids in the polypeptide chains. Understanding the rules that govern how these simple lines of amino acids fold – i.e., reading between the lines – is a key goal in modern biology.

Indeed, such linear polypeptides self-assemble into a wide array of elegant 3D protein structures, with great variation in size and in the size of their substrates. Proteins can adopt monomeric or oligomeric structures, as well as forming elaborate higher order multi-protein complexes. Beneath this complexity observed across protein structures (e.g., currently 62957 protein structures populate the PDB (2)), there is an underlying simplicity: 1393 protein folds are known (using the SCOP classification (3)), eight possible secondary structural elements (4), with an underlying alphabet of 20 amino acids. The primary sequences of proteins are somewhat more complex than that of DNA, with only four bases, but considerably less complex than the primary sequences of carbohydrates, with their ability to branch in many ways (5). Proteins are large molecules, folding into sufficiently rigid scaffolds to perform their function effectively; with a low enough surface area to volume ratio to prevent aggregation in the crowded cell; and with a suitably sculpted surface to direct substrate to its active site (6), the latter often being the deepest cavity in the protein (7).

In the study of Mittal et al. (1), analysis of a set of nearly 4000 folded protein crystal structures finds a rather narrow range of variation in amino acid composition, suggesting that this specific ratio of amino acids is a prerequisite for a successfully folded protein (Table 1 of reference (1)). As a control, the authors observe a different stoichiometry for proteins which populate significantly unfolded states (Figure 5B of reference (1)). Intriguingly, for the stoichiometry of the folded protein set, the highest abundances are observed for Ala, Val, Leu, Gly and Glu, with the lowest for Met, Cys, Trp and His. Associated with these averages are standard deviations across the set of ~4000 crystal structures which span 1.0–3.4%. This range is colourfully referred to by the authors as the “margin of life”.

The authors draw an analogy between their work and that of Erwin Chargaff, who demonstrated via elegant experiments the 1:1 stoichiometry of G:C as well as A:T in DNA. Chargaff also used the four nucleotide bases to deduce that the composition of DNA is organism-specific, showing their relative proportions varied between organisms. The latter rather simple observation hides the tremendously intricate differences in genomic sequence that define species. In the same way,
the “margin of life” though apparently narrow must be sufficiently broad, and the ordering of amino acids within the confines of this stoichiometric range sufficiently flexible, to permit the complexity in the folded shape and behaviour of functional proteins.

The underlying physics which dictate protein folding is the subject of much debate, as Mittal et al. acknowledge in their introduction (1). The authors examine the average radial distribution of the 20 amino acid types around a given type of amino acid, as measured by C$_\alpha$C$_\alpha$ distances. From their analysis of the extent of these contacts and the similarity in profile of their distribution for a given amino acid type, the authors conclude that protein folding is dictated simply by frequencies of occurrences of individual amino acids. This leads the authors to highlight solvation and packing effects as the key guiding influences in the protein folding process.

Interestingly, the authors’ analysis of the radial C$_\alpha$C$_\alpha$ contacts in crystal structures bears some resemblance to the derivation of knowledge-based potentials used in protein folding prediction (8), although the latter are normalized with respect to a hypothetical reference state, e.g. frequency of contacts expected from random mixing of amino acids and solvent. Despite being low in resolution, the latter potentials have shown some success in predicting 3D structures of small proteins but typically require augmentation with other terms and subsequent force field refinement in order to yield results accurate to a resolution of ~1.5 Å (9-13). Of course, we note that knowledge-based potentials and the C$_\alpha$C$_\alpha$ neighbourhoods calculated in (1) involve spherical averaging, whereas proteins are inherently asymmetric molecules, a key to understanding their biological function.

Nevertheless, the insights derived from the work of Mittal et al. represent an important contribution to the debate surrounding the basis of protein folding; extension of this analysis to a greater number of the known protein structures, as well as to detailed examination of sub-sets such as structural proteins, will provide further understanding of how linear chains of amino acids adopt their exquisite 3D structures.

References