“Chargaff’s Rules” for Protein Folding: Stoichiometric Leitmotif Made Visible

http://www.jbsdonline.com

“Everything should be as simple as it is, but not simpler!”
-Albert Einstein

Protein folding! The first thing that almost always comes to mind when someone hears this term is Anfinsen’s hypothesis. So much so that protein folding and Anfinsen’s hypothesis have long since been considered synonyms of each other.

Anfinsen’s hypothesis laid the ground rule for protein folding. The rule does not even need rote learning, rather common sense: sequence determines structure which, in turn, determines function.

Then arrived several models and theories which tried to explain exactly how a protein folds into a perfect functional form commencing as soon as the amino-terminal chain emerges from the ribosome or after being fully synthesized. The models were all woven around a basic set of rules: no new bond making or breaking in the folded protein and global minimum-most free energy for the native structure. None of the methods, however, provide adequate instructions for the exact folding process. This can be seen by the simple fact that we are still far away from predicting an accurate, valid three-dimensional structure right from an arbitrary sequence using computational tools, with absolutely no mediator structures in-between. It is largely known that in vivo, the folding process for some proteins requires the aid of molecular chaperones.

In the manifold attempts towards elucidating the real mechanism, a recently published report by Mittal et al. (1) comes as a fresh breeze with a novel perspective. In accordance with Einstein’s remarks above, they report a ‘surprisingly simple unifying principle’ in that the protein folding appears to be dictated by ‘a narrow band of stoichiometric occurrences of amino-acids in primary sequences, regardless of the size and the fold of a protein’. The authors further maintain that contrary to all known perceptions, ‘preferential interactions’ between amino acids do not drive protein folding.

The authors devised an elegant and simple way of approaching the problem. Using the crystal structures of 3718 proteins present in Protein Data Bank (PDB), they analyzed the backbones in terms of the neighbourhood of every Cα atom. This was based on the assumption that if two amino acids interact via side chains, their respective Cα atoms would be ‘neighbours’ fixed in space, disregarding their actual position along the protein sequence. The Cα atoms occurring in the protein’s ‘center’ would be expected to be surrounded by a higher number of Cα atoms of other amino acids. Considering each of the amino acids individually, a

Seema Mishra
National Institute of Biologicals
Ministry of Health and Family Welfare
A-32, Sector 62, NOIDA, U. P.,
India 201307

Corresponding Author:
Seema Mishra
E-mail: seema_nib@yahoo.com
20 x 20 matrix of the number of neighbours within a defined neighbourhood distance was reported for each protein. Plots for the number of times a specific amino acid appeared as a neighbour within a fixed distance were derived to investigate the presence of any ‘preferential interactions’. Here, it should be noted that the non-covalent interactions between amino acids are the basis of the development of knowledge-based potentials, which in turn form the underpinning of protein structure prediction by modeling and simulation (2, 3 and references therein). The authors investigated whether amino acids prefer certain neighbourhoods arising out of non-covalent electrostatic, hydrogen bonds, hydrophobic and van der Waals interactions.

A sigmoidal trend was observed for the neighbourhoods of all the 20 amino acids and was found to be independent of the nature and identity of an amino acid. The immediate implications were that either several spatial distributions exist individually or there is one unifying spatial distribution pattern. A generalized, single equation fitting all the sigmoid was found and it brought forth the latter implication in full clarity. Further, a simple reasoning was applied that assuming there are no ‘preferential interactions’ involved, an amino acid that occurs as the highest number of times in the primary sequence would be expected to occur as a neighbour of all the 20 amino acids, including itself, the highest number of times. As expected, good correlation was established between the total number of contacts made by an amino acid and its stoichiometry (frequency of occurrence) in folded proteins. By direct implication, for the unstructured proteins, the authors found the stoichiometry to deviate from that of amino acids in folded proteins.

These studies preclude ‘preferential interactions’ and define stoichiometry of amino acids in the primary sequence governing the folding process. The very fact that the rules governing the folding of these proteins are universal makes it all the more inspiring.

The protein folding thesis by Mittal et al. is further strengthened by recent external evidence (4). Several protein properties, that range from size to folding cooperativity to dispersions in melting temperature, were found to depend universally on the number of residues, N. Even the timescale to reach the ‘native basin of attraction’ (native state) can be estimated in terms of ‘N’ residues. These predictions have been well-supported by experiments. But there is one major difference here, that it is the stoichiometry of amino acids rather than their number that has been found to govern.

The field of protein folding has been so intriguing that the ideas to understand the basic tenets have been sought even from the laymen. An online game, Foldit (5), recently-developed, serves to further this cause.

Until the 1990’s, protein folding was, in terms of computation, an impregnable entity and has gradually been gaining ground since. One might be over-optimistic in that, given this universal principle based on stoichiometry, biologically meaningful ab initio modelling will indeed be possible in near future. An entirely ‘new’ structural space of many hypothetical as well as trans-membrane proteins will be visible, ‘new’ here referring to the fact that the basic premise will be applicable to one and all of the proteins and hence that, ‘there is nothing new under the sun!’. With this, widely acknowledged Darwinian evolution rules might be re-written with some clauses added and others deleted.

It is well-known that a structure is more conserved than a protein’s sequence. Mittal et al. have taken backbone atoms to investigate the folding process which clearly leads to universal applicability. It would mean that the basic fold is the same across the protein kingdom. Specificity and uniqueness are, but, imparted by the side chains in proper rotameric state. This conformational specificity is achieved as a result of packing and many weak interactions (6). Hence, in order to achieve a proper, functional and unique structure, stoichiometry does appear to play a role, albeit not per se.

While these findings are based entirely on theoretical approach, it is essential to consider these in the light of real scenario. Proper folding is usually thought to occur after a part or a domain of the protein is translated by the ribosomes. Apart from molecular chaperones to aid, existence of RNA-mediated chaperone type for folding de novo is documented (7). Co-translational kinetics may also play a role (8).

Protein folding in vivo is subjected to a different environment than in vitro. While in vivo, a cramped environment is provided to the protein; in vitro, it is generally a much more relaxed one. The cramped environment of the cell may force the protein to misfold causing aggregation (9). It is equally plausible that given relatively less number of solvent molecules due to the presence of more solutes around in vivo, the excluded volume effects have to be taken into consideration by the protein and fold correctly. Mittal et al. propose that folding has to be done while considering the structural constraints of the amino acid sequence and the order in which a given stoichiometry appears. Further, the authors opine that another parameter for folding is the individual shape characteristics of amino-acids along the sequence minimizing the surface-to-volume ratio. Entropic and energetic considerations of the folding in this manner would also seem favourable. When constituent molecules are used up in stoichiometric ratios as an organized principle, it is logical that the ordered folding may proceed at the expense of increased entropy because of the disorder induced by the unfolded ensembles in the surroundings.
The stoichiometric ratios for each amino acid in folded proteins give rise to “Chargaff’s rules” for protein folding, and these rules suggest that leucine has the greatest average percentage occurrence (9.0 ± 2.9) and tryptophan (1.3 ± 1.0) and cysteine (1.8 ± 1.5) the least. This seems plausible since leucine, being a hydrophobic residue and, being among the most abundant in proteins, is expected to play a major role in the folding or packing arrangements of cytosolic proteins that tend to fold in a compact manner away from the solvent as well as the folding of trans-membrane proteins in a hydrophobic environment. No wonder that leucine zippers of transcriptional activators are an ideal system to study the folding dynamics (10). However, this is not to assume that tryptophan and cysteine residues have little role in the folding process because of their lowest percentage occurrence. On the contrary, these residues have been shown to be necessary for proper folding of neprilysin (NEP)/endothelin-converting enzyme (ECE) family of metalloproteases, to cite one example (11). Further, if the folding process is uniform for both, intuitively, the question is: what exactly distinguishes the cytosolic and trans-membrane proteins with regard to their structure and hence, specific fate?

If we take this mechanism to be true, then it is entirely explainable that chaperones, obviously, have an indirect, supportive role in either coordinating or overseeing the process. It appears that along the path to a proper fold, if there were mediator structures mediating between any two flanking native-like conformations attempting to generate a consensus, they might force decisions into one direction that may not generate a proper fold. So, if a protein is to get its desired structure and function, the trick is to let it fold naturally without the help of any mediators; and it is here that the inter-play of stoichiometric ratios of amino acids driving the folding is widely apparent. The question remains, how disrupting a part of the protein by mutations, resulting in insertion, deletion or substitution of amino acids and hence, disruption of stoichiometric occurrences, is expected to let the protein be viable in its remaining fold as is sometimes observed.

Protein misfolding may also be well-explained in the context of this new finding. Other proteins within a milieu may compete for specific amino acids thereby leading to misfolding of those deprived of the amino acids in required stoichiometric ratios.

Do the proteins prefer just the native fold? Or variations leading to an alternate functional folded state are plausible? Two proteins were designed with 88% sequence identity and their different structures and functions were found to be encoded by only 12% of amino acids (12). It was further suggested that the free energy change of alternative folded states may be much more favourable. Depending upon stoichiometry and thermodynamics, the alternative folded states may be formed and preferred over the native.

Needless to say, common sense dictates that protein folding is environment- or context-dependent and therefore, folding may not be dependent upon stoichiometry alone. Many proteins are known to fold only in the presence of a ligand. Paralogs, common examples of which are myoglobin and haemoglobin with poor sequence identity but similar secondary structures, reside in different cellular micro-environments. Some proteins such as haemoglobin function only in a multi-domain form. If we consider a particular bacterium which has to live in an environment where some amino acids are limiting, there must be some mechanism for ensuring the availability and hence, maintenance of adequate stoichiometry of amino acids.

Due consideration should also be given to the possibility of the effect of codon bias in different organisms. In the case of silent mutations, another amino acid having the same physico-chemical properties is used to form a similar fold with virtually no phenotypic consequences. Providing an exception to Anfinsen’s dogma, a silent polymorphism due to a single-nucleotide change that altered P-glycoprotein’s function, was determined to be resulting from a fold different from the wild type (13). A rare codon for isoleucine was substituted in the place of wild type codon. In such a situation, exactly how the stoichiometric occurrence of the amino acids encoded by rare and/or cognate codons is expected to drive the correct fold-formation is worth exploring further.

In disulfide-rich proteins, disulfide bond formation is found linked to protein folding activity thermodynamically. Structural disulfide bonds are formed only when there is correct backbone-to-backbone distance and proper orientation of cysteine residues. A mean distance of 5.6 Å is required between two alpha-carbons of cysteine residues of adjacent peptide backbones (14). In the case of oxidative folding of antibody Fab fragment, formation and re-arrangement of disulfide bonds is involved, and so, enzymes such as protein disulfide isomerase need to enter into the picture. Thus, it is apparent that it is not just stoichiometry (two cysteine molecules for one disulfide bond formation) that plays a role in oxidative protein-folding process, enzymes are essential here in the larger scheme of things.

The studies by Mittal et al., take into consideration only the proteins with more than 50 amino acid residues. It will be worthwhile to explore the folding of small proteins and peptides that are also crucial to the proper functioning of the cell in light of the new findings. For example, 15-amino acids-long helper T lymphocyte epitopes or the nonameric cytotoxic T lymphocyte (CTL) epitopes need to contact both major histocompatibility complex (MHC) and T cell receptors (TCR).
at any given point of time, and so, how do they adopt their particular fold? An idea about the involvement of stoichiometry in their folding is supported by a theoretical instance (15). In this study, most of the CTL epitopes shown to be interacting with MHC molecules have leucine or valine at the two primary anchor positions and also at the positions where TCR is expected to make a contact. Other examples of peptides for possible further studies are the five-amino acids-long met-enkephalin and the nine-amino acids-long hormone vasopressin.

Rather uniquely, the “Chargaff’s rule” for protein folding reported by Mittal et al. simplifies both the gross and subtle mechanisms of protein folding and paves novel pathways for many more discoveries, yet to be unravelled. The immediate medical implication seems to be, obviously, a thorough understanding of protein misfolding and aggregation in diabetes, cancer, Alzheimer’s and Parkinson’s diseases among others ultimately leading towards the discovery and development of successful drugs. A recently developed highly effective prodrg with enhanced stability and long-term bioavailability, supramolecular insulin assembly (SIA-II) of insulin oligomers for diabetes treatment, has been developed taking into account the notions of protein folding (16). And what is more, productive and meaningful ab initio protein structure prediction is in our hands now—well, almost! Interestingly, these findings by Mittal et al. have been derived using standard mathematical principles for a biological problem. Who knows, there might be a magical number, similar to the universal ‘Pi’ (π) or the golden mean ‘Phi’ (Φ) presiding over many natural phenomena, that governs even the protein folding process!

References