Preface
PREFACE

Proteins are covalently linked linear chains of amino acids. Protein chains adopt unique three-dimensional (3-D) structures known as ‘native structures’ that are perfectly adapted to the carrying out of their intended biological functions. All the information required for specifying the native 3-D structure of a protein is contained within its amino acid sequence, and this is the only information that passes in the form of gene sequences between generations of cells, or organisms. Of course, protein chaperones and other folding aides [such as peptidyl prolyl cis-trans isomerase (PPI) and protein disulfide isomerase (PDI)] are important too for the folding reaction to occur properly [1-4], in that chaperones improve the overall yields of folding by reducing inter-molecular associations and preventing irreversible aggregation, whereas molecules such as PDI, and PPI, actually catalyse folding without however necessarily effecting dramatic improvements in folding yields; even so, neither class of helpers has yet been shown to help the amino acid sequence in determining the native three-dimensional structure. Rather, many proteins demonstrably refold to native structure from a denatured state in vitro in the absence of folding aides, establishing that the covalent chemical structure of a protein chain (i.e., the amino acid sequence) is alone capable of determining its 3-D structure, with aides merely helping to improve folding rates and yields, but not the structure resulting from the folding reaction.

How simple or complicated is the dictation of 3-D structure by an amino acid sequence? Polypeptides have a very large number of degrees of rotational freedom. Every bond in a protein molecule that is a single, or ‘sigma’, bond can conceivably allow rotation of entire groups of atoms flanking it. Thus, assuming a conservative estimate of 3 rotameric states per sigma bond, and an equally conservative estimate of 3-4 four rotatable sigma bonds per residue, every residue in a protein chain can be reckoned to be capable of adopting any one of about 10 different 3-D structures at any point of time. Thus, even with such conservative estimates, a small protein of 100 residues can be reckoned to be capable of adopting any one (unique, native) structure from amongst an astronomically large population of $10^{100}$ conceivable structures that comprise its entire ‘conformational space’. How does the amino acid sequence
provide information to the chain regarding which single, unique structure from amongst this vast universe of structures is required to be adopted? If a polypeptide were required to sample a majority of this large set of structures before settling into its native structure, through a conformational search driven by thermodynamics considerations alone (i.e., by the ‘thermodynamic control’ model) [5], clearly it would require more time to do so than is available in the rest of the universe’s predicted lifetime. Yet, as is well known, polypeptides can fold spontaneously over timescales as short as microseconds, milliseconds or seconds, establishing that the chain cannot possibly be accessing even a significant fraction of its conformational space (the well known ‘Levinthal’s paradox’) [6-8]. Evidently, therefore, the information content in a protein’s amino acid sequence must comprise of not just the final structure required to be adopted but also the mechanism by which that structure is adopted. If, for example, small sets of interactions occur in step-wise fashion to progressively restrict the chain’s conformational freedom during folding, such that large portions of the protein’s conformational space become kinetically inaccessible to the folding chain which is effectively ‘forced’ down certain designated pathways of folding leading rapidly to native structure (the ‘kinetic control’ model) [9,10], clearly this information too must be encoded within the amino acid sequence. In certain instances, despite the potential for orderly and progressive formation/assembly of sub-structures, energy barriers to certain stages of folding and assembly - comprising of high energy transition states - might conceivably exist to slow down chain folding sufficiently to allow individual molecules adopting any one of multiple available schemes/mechanisms/pathways of folding within a population, to globally optimize residue-residue interactions before the chain’s final descent into a native structure [the ‘non-two-state’ folding scenario] [11,12]. The amino acid sequence must thus encode all options for the adoption of multiple pathways of folding, as well as information concerning the precise definitions and constitutions of key interactions defining transition states and residues constituting folding nuclei, so that the same may be brought together at appropriate times during folding.

As native structures are clearly adopted within finite time without providing the chain with adequate scope to attain true thermodynamic equilibrium through random exploration of its conformational space [both because there isn’t enough time available, and because high energy barriers, and early interactions prevent the chain’s
access to many low-energy conformations that are never sampled] the possibility remains alive that the native structure does not always necessarily correspond to the most stable of all conceivable conformations. If so, clearly the amino acid sequence is also required to encode information ensuring sufficient ‘kinetic stability’ within the molecule to ensure that the native structure survives for as long as it is required to, in order to perform the task(s) required. In addition to such kinetic stability, sufficient thermodynamic stability must also be imparted to the native structure to allow both formation and turnover of molecules; in most globular proteins, one finds that the free energy difference between native and unfolded states is of the order of 5-15 kcal/mol, or the equivalent of 3-4 hydrogen bonding interactions, signaling that the thermodynamic stability of the native state results from a small difference between large numbers of stabilizing (hydrogen-bonding, hydrophobic and electrostatic) interactions and destabilizing interactions (principally van der Waals interactions of the steric kind).

In some proteins, local interactions are seen to dictate formation of secondary structures that then condense together (the ‘hierarchical assembly’ model) [13,14] whereas in other proteins, an early compaction of the chain (through a ‘hydrophobic collapse’) [15] facilitates long-distance interactions that then restrict the scope for secondary structure formation through local hydrogen bonding interactions. In both of these modes, the scope remains for ‘diffusion-collision’[16,17] based interactions of the local as well as long-distance kinds, and also for ‘nucleation-condensation’ [18] reactions involving nucleation-propagation of helices and beta turn elements and condensation with other secondary structures.

All of the above kinds of structure-related information is required to be present within the amino acid sequence, which thus ends up being the principal repository of information determining: (i) the final structure intended to be adopted from amongst an astronomically large number of conceivable structures, (ii) the manner in which the structure is required to be adopted, while avoiding formation of alternative structures, (ii) the thermodynamic stability of the structure in relation to those of alternative structures, which would ultimately determine whether the system is ‘stable’ or ‘metastable’, (iv) the kinetic stability of the structure, which would determine the survival of the structure, in concert with (v) the manner in which the molecule would
be turned over in vivo through proteolytic mechanisms. In other words, every kind of 3-D structural information is encoded within the amino acid sequence.

And yet we understand very little about the nature of coding of this information. To take just two issues that are central to the experimental work presented in this thesis:

(I) We do not yet understand the role of the polarity of the polypeptide backbone in determining the folded state. Individual amino acids have an amino end and a carboxyl end. When they are bonded together into a polypeptide too, therefore, the backbone of the polypeptide ends up having a direction, with one end containing the amino terminus and the other end, the carboxyl terminus. In principle, therefore, two chains of identical residue sequence could thus have backbones running through them in exactly opposite directions. Would such chains fold with equal facility? Would their adopted structures have any geometrical relationship to each other?

(II) We do not yet understand whether folding information is encoded predominantly locally, or globally, within the chain. If encoding were strictly global, major alterations of sequence effected e.g., through ‘shuffling’ of large tracts of sequence encoding individual secondary structural elements, or assemblies of such elements, could be expected to disturb the folding reaction sufficiently to prevent folding. Going by the well known ‘two-state’ model of folding, a polypeptide chain corresponding to a naturally-occurring protein can exist in just two states: the folded (native) state, and the unfolded state, with no other states being allowed. Thus major alterations of sequence disturbing the encoding of information would be expected to give rise to unstructured states that either remain soluble or precipitate out of aqueous solution due to loss of solubility, or due to aggregation occurring through non-sequence-specific hydrophobic and hydrogen-bonding interactions. On the other hand, if encoding were reasonably local, such shuffling would give rise to structured (even if aggregation-prone) non-native states, providing glimpses into whether folding occurs hierarchically through the formation and assembly of substructures or globally through a two-state, all-or-none sort of collapse.
Besides the obvious intellectual satisfaction of addressing such questions concerning the sequence-structure relationship in proteins as stated above, several useful applications too could result from such studies. Anticipating a protein's fold represents a classic challenge; since nature knows how to decode sequence information to create folded states with high fidelity and reproducibility, it is tempting to postulate that computers too might one day be persuaded to provide structures from sequence information alone, either through empirical routes, e.g., through comparisons of existing sequences and structures, or through hard simulation of the folding process. Still, the complexity of modeling interactions \textit{ab initio} is so high that such attempts can never succeed without large amounts of experimental inputs which can only come from the imaginative asking and addressing of fundamental questions. Every newly sequenced gene gives rise upon translation into a naturally-occurring amino acid sequence that 'knows' how to fold. In the era of genomics that has produced such a profusion of amino acid sequences, therefore, it is more urgently required than ever before to understand the sequence-structure relationship in proteins, to aid computational efforts at predicting or redesigning structure, since clearly experimental methods of determining structure cannot keep pace with the rate at which sequence information is becoming available.
References


