J. Mol. Biol. (1976) **104**, 59–107

A Simplified Representation of Protein Conformations for Rapid Simulation of Protein Folding

MICHAEL LEVITT

Weizmann Institute of Science, Rehovoth, Israel
and
Medical Research Council Laboratory of Molecular Biology
Hills Road, Cambridge, England†

(Received 12 December 1975, and in revised form 19 February 1976)

This report is one of a series of papers that introduce and use a new and highly simplified treatment of protein conformations. The first paper (Levitt & Warshel, 1975) outlined the approach and showed how it could be used to simulate the "renaturation" of a small protein. The present paper describes the representation in some detail and tests the methods extensively under a variety of different conditions. The third paper (Warshel & Levitt, 1976) is devoted to a study of the folding pathway and stability of a mainly α-helical protein.

In this work, I show how the concept of time-averaged forces, introduced previously (Levitt & Warshel, 1975), can be used to simplify conformational energy calculations on globular proteins. A detailed description is given of the simplified molecular geometry, the parameterization of suitable force fields, the best energy minimization procedure, and the techniques for escaping from local minima. Extensive tests of the method on the native conformation of pancreatic trypsin inhibitor show that the simplifications work well in representing the stable native conformation of this globular protein. Further tests show that simulated folding of pancreatic trypsin inhibitor from open chain conformations gives compact calculated conformations that have many features in common with the actual native conformation. Folding simulations are done under a variety of conditions, and the relevance of such calculations to the actual *in vitro* folding process is discussed at some length. These same techniques have many potential applications including enzyme–substrate binding, changes in protein tertiary and quaternary structure, and protein–protein interactions.

1. Introduction

Protein molecules fulfil almost all the catalytic and structural roles in the living cell; they are both the machine tools and building blocks of the cell's factories. Such functional and structural versatility is entirely due to the folding of different amino acid sequences into different three-dimensional conformations. In each of these folded conformations, the position of every atom is precise and depends uniquely on the particular amino acid sequence (Anfinsen *et al.*, 1961; Anfinsen, 1973). This relationship between a protein's sequence and its three-dimensional structure constitutes the second part of the translation of genetic information into functional

† Present address.

protein molecules. The first part, the synthesis of the correct sequence of amino acids from the DNA sequence, differs in almost all respects from the second part: protein synthesis is a complicated biochemical process depending on hundreds of enzymes, transfer RNAs, and the ribosome; protein folding is a simple physical process depending on the same interatomic forces that stabilize the simplest molecules. Besides the attractiveness to theoreticians of the protein folding problem, the ability to calculate the conformation of any protein from its sequence would be an invaluable aid to molecular biology in general.

In principle, one should be able to use the methods of conformational analysis developed for small molecules on proteins. With these methods, one finds stable conformations of the molecule by changing selected variables to minimize the total energy, which is expressed as an analytical function of the atomic positions, the chemical connectivity, and the interatomic forces. In practice, severe problems arise when extending techniques that work well for small systems to much larger systems. (a) Proteins have too many atoms. As an important contribution to the energy is summed over all pairs of atoms, calculating the total energy of a protein is time-consuming. (b) Proteins are stable in water at room temperature. While the properties of small molecules *in vacuo* at low temperatures can be computed fairly easily, much less is known about the effect of the solvent and atomic thermal motion on the interatomic forces. (c) Protein structures need to be described by too many variables. Even if one considers only the torsion angles about single bonds, a small protein still has several hundred degrees of freedom, making energy minimization much less efficient.

Previous theoretical studies on proteins illustrate these difficulties clearly. Much attention has been given to the allowed conformations of single amino acids, essentially a small-molecule problem. The earliest work was based on the simple idea of forbidding conformations that have very close non-bonded contacts (Ramachandran *et al.*, 1963), whereas the most recent studies include solvent and entropic effects (Lewis *et al.*, 1973). Studies of bigger systems with up to, say, 20 residues have led to the introduction of more powerful techniques (Gibson & Scheraga, 1967,1969), but the results were disappointing as in calculations on gramicidin S (Vanderkooi *et al.*, 1966; Liquori *et al.*, 1966; Momany *et al.*, 1969; De Santis & Liquori, 1971). In recent years, following the first energy calculations on known protein conformations (Levitt & Lifson, 1969), more attention has been given both to the energy refinement of X-ray co-ordinates of proteins (Levitt, 1974*b*; Warme & Scheraga, 1974; Hermans & McQueen, 1974; Gelin & Karplus, 1975), and to the binding of a substrate to a protein (Platzer *et al.*, 1972; Levitt, 1972,1974*b*). In all these cases the calculated changes in conformation are small (<1 Å).

Recently, Burgess & Scheraga (1975) applied the methods used before, in the refinement of protein X-ray co-ordinates, to a conformation of pancreatic trypsin inhibitor that had the correct local structure (all (ϕ, ψ) angles had been set to within 30° of the native values), but did not have the correct native tertiary structure. Although this calculation required considerable computer time, their results were disappointing: it was not possible to refine the (ϕ, ψ) angles to get back the native tertiary structure, even if the S—S bonds were artificially brought together. These difficulties arise from the intrinsic complexity of protein structure in the conventional all-atom representation, the time-consuming evaluation of the energy with so many atoms, and the large number of variables that must be considered.