Modelling of peptide and protein structures*

Review Article

S. Fraga¹ and J. M. R. Parker²

¹ Department of Chemistry and ² Alberta Peptide Institute and Department of Biochemistry, University of Alberta, Edmonton, Alberta, Canada

Accepted August 12, 1993

Summary. The modelling of protein structures (whether isolated, in solution, or involved in recognition processes) is reviewed, free of any mathematical apparatus, to provide an overview of the concepts as well as leading references. A general feeling for this field of work is first established by a sampling of some impressions on its difficulties and chances of success. Then, the main body of this work examines the information available (databases and parameters), presents the theoretical foundations for the modelling procedures (with emphasis on the potential energy functions), surveys the existing simulation techniques and prediction methods, and discusses the problems still to be faced. For completeness, a representative list of existing software packages is presented in the Appendix.

Keywords: Amino acids – Peptides – Proteins – Modelling – Simulations

Introduction

The a priori prediction of peptide and protein structures will have far-reaching implications in academic as well as biotechnology research. Recently, this challenging research problem has been the centre of considerable attention. Success in this endeavour will complete the description of the genetic code and will have a social as well as an economic impact, as it will open the way for the development of new drugs, synthetic vaccines, and industrial enzymes. Protein engineering will endeavour to design new proteins or to change the structural and/or functional characteristics of existing peptides and proteins for specific purposes (van Gunsteren, 1988).

A staggering amount of original work has been carried out, resulting in a wealth of information and establishing most of the necessary formulations, which have been incorporated in a variety of software packages. The use of such packages as black boxes has expanded dramatically and it is now appropriate to present a conceptual review of the field. This information may help experimental researchers, interested in complementing, expanding or interpreting their work with simulations, in approaching such a task with realistic expectations.

A feeling for the state of this field may be obtained from a sampling of the impressions of active researchers. The advent of fast, large-capacity computers induced an early euphoria ['Many are now racing to see who can be the first to calculate correctly the three-dimensional structure of an enzyme from its sequence' (Bradley, 1970)] but the difficulties soon became evident ['Even very short peptides (5–30 amino acids) are difficult to model accurately in the absence of structural data. There are simply too many degrees of freedom' (Wilson and Klausner, 1984), to the point that the goal almost seemed unattainable ['The prediction of the three-dimensional structure of a protein from its amino acid sequence remains one of the fundamental unsolved problems in molecular biology' (Thornton, 1988)]. Many will be the rewards of solving this problem ['Deciphering the rules through which amino acid sequences determine protein folding will be a major scientific and technological advance' (King, 1989)] and therefore one should reexamine the fundamental considerations when tackling it ['The computational task of protein structure prediction is believed to require exponential time, but previous arguments as to its intractability have taken into account only the size of a protein's conformational space. Such arguments do not rule out the possible existence of an algorithm, more selective than exhaustive search, that is efficient and exact' (Ngo and Marks, 1992)]. Work nevertheless continues and progress is made, and although doubts still linger and refinements must be made ['It remains to see whether some of the apparent advances made in the past few years are real or illusory and whether the protein folding problem is well on its way to being solved. One problem that clearly remains is the elucidation of the magnitude of the various forces that determine the delicate free energy balance between the folded and unfolded states of the protein (Honig et al., 1993)], the problem will perhaps be finally solved ['As is evident in the many efforts to characterize the structures and stabilities of various folding intermediates and their roles in folding reactions, it seems very likely that substantial progress will be made in solving the folding problem in the next few years. The development of a folding code, analogous to the genetic code, will allow us to expand the central dogma of molecular biology to the prediction of the 3-D structure adopted by a given amino acid sequence' (Lecomte and Mathews, 1993)].

The description presented below, with a review of the various components in protein modelling, will confirm the above hopeful, but realistic, expectations. No effort has been made in presenting a complete survey of all the existing literature, but we are confident that the most important references are included or may be traced back from the ones given in this work. The reader entering this field for the first time should, first of all, get acquainted with the work of Richardson (1981), Jaenicke (1987), and Chothia (1990), describing folding and associations in proteins.