Lecture Notes in Chemistry 66

S. Fraga J. M. R. Parker J. M. Pocock

Computer Simulations of Protein Structures and Interactions

Springer

S. Fraga J. M. R. Parker J. M. Pocock

Computer Simulations of Protein Structures and Interactions

Springer

Authors

S. Fraga
Department of Chemistry, University of Alberta, Edmonton, AB, Canada T6G 2G2
and
Departamento de Química Física Aplicada, Universidad Autónoma de Madrid
28049 Canto Blanco (Madrid), Spain

J. M. R. Parker
Alberta Peptide Institute and Department of Biochemistry, University of Alberta
Edmontom, AB, Canada T6G 2S2

J. M. Pocock
Department of Biochemistry, University of Dundee, Dundee, Scotland DD1 4HN
United Kingdom

Cataloging-in-Publication Data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Fraga, Serafin:
Computer simulations of protein structures and interactions /
S. Fraga ; J. M. R. Parker ; J. M. Pocock. - Berlin ; Heidelberg
; New York ; London ; Paris ; Tokyo ; Hong Kong ; Barcelona
; Budapest : Springer, 1995
(Lecture notes in chemistry ; 66)
ISBN 3-540-60133-3 (Berlin)
ISBN 0-387-60133-3 (New York)
NE: Parker, J. M. Robert:; Pocock, Jennifer M.:; GT

ISBN 3-540-60133-3 Springer-Verlag Berlin Heidelberg New York

Typesetting: Camera ready by author
SPIN: 10500620 51/3142 - 543210 - Printed on acid-free paper

Preface

The *de novo* prediction of the tertiary structure of peptides and proteins and, ultimately, the development of molecular switches, synthetic vaccines, pepzymes, peptidomimetics, ..., are the goals of research in biotechnology. This field relies on a cooperative effort between many branches of the life and natural sciences and this work strives to present the corresponding picture, from quantum mechanics to rational drug design, through computer science, synthetic chemistry, immunology, pharmacology, ... We hope that it will serve its dual purpose, as a learning tool and as reference.

In this type of work it is advisable to try and profit from the knowledge of experts in the various fields and we have been fortunate in counting with the comments of G. Arteca (Laurentian University), A. Cameron (Synphar Laboratories Inc.), J.M. Garcia de la Vega (Universidad Autonoma de Madrid), M. Klobukowski (University of Alberta), G. Kotovych (University of Alberta), G.R. Loppnow (University of Alberta), I. Rozas (Instituto de Química Médica, C.S.I.C.), D.S. Wishart (University of Alberta), H. Yamamoto (Osaka National Research Institute), and S. Yoshikawa (Osaka National Research Institute).

And, as always, we acknowledge the efficiency, patience, and initiative of J. Jorgensen (Department of Chemistry, University of Alberta).

S. Fraga, J.M.R. Parker, J.M. Pocock
Dundee Edmonton Madrid

Acknowledgments

In this work, data, methods and formulations, and results from many sources are examined, discussed, and compared, with reproduction in some instances of copyright material.

We would like to acknowledge here our indebtedness to those organizations from whose publications information and/or copyright material has been used:

Academic Press Limited, Adenine Press, Alan R. Liss, Inc., American Association for the Advancement of Science, Annual Reviews Inc., Bailliére Tindall, Biophysical Society, Butterworth-Heineman, Cambridge University Press, Consejo Superior de Investigaciones Científicas (Spain), Current Biology Ltd., Elsevier Biomedical Press, Elsevier Science Publishers B.V., ESCOM Science Publishers B.V., European Peptide Society, International League Against Epilepsy, International Union of Crystallography, Interscience Publishers, IRL Press, John Wiley & Sons, Inc., MacMillan Journals Ltd., Munksgaard Ltd., National Academy of Sciences (U.S.A.), National Research Council (Canada), New Forum Press, Inc., North-Holland Publishing Company, Oxford University Press, Pergamon Press, Inc., Pharmacotherapy Pub. Inc., Plenum Publishing Corporation, Portland Press Ltd., Protein Research Foundation (Japan), Raven Press, Ltd., Real Sociedad Española de Química, Societa Chimica Italiana, Springer-Verlag, The American Chemical Society, The American Institute of Physics, The American Physical Society, The American Society for Biochemistry and Molecular Biology, Inc., The Chemical Society (U.K.), The National Academy of Sciences (U.S.A.), The Physical Society (Japan), The Protein Society, VCH Verlagsgesellschaft mbH, Wiley-Liss, Inc.

Specific mention must be made, in particular, of the following works from which copyright material has been taken:

A.D. Buckingham. *Permanent and Induced Molecular Moments and Long-Range Intermolecular Forces*, in *Intermolecular Forces*, edited by J.O. Hirschfelder. Advances in Chemical Physics, vol. 12, pp.107-142. Interscience Publishers, New York (1967).

D.A. Clark, G.J. Barton, and C.J. Rawlings. *A Knowledge-Based Architecture for Protein Sequence Analysis and Structure Predictions*. Journal of Molecular Graphics *8*, 94-107 (1990).

A. Godzik, A. Kolinski, and J. Skolnick. *De Novo and Inverse Folding Predictions of Protein Structure and Dynamics*. Journal of Computer-Aided Molecular Design *7*, 397-438 (1993).

R.D. King and M.J.E. Sternberg. *Machine Learning Approach for the Prediction of Protein Secondary Structure*. Journal of Molecular Biology *216*, 441-457 (1990).

G.H. Loew, H.O. Villar, and I. Alkorta. *Strategies for Indirect Computer-Aided Drug Design*. Pharmaceutical Research *10*, 475-486 (1993).

M.A. Navia and D.A. Peattie. *Structure-based Drug Design: Applications in Immunopharmacology and Immunosuppression*. Immunology Today *14*, 296-302 (1993).

W.R. Taylor. *Protein Structure Prediction*, in *Nucleic Acid and Protein Sequence Analysis. A Practical Approach*, edited by M.J. Bishop and C.J. Rawlings, pp.285-322. IRL Press, Oxford (1987).

Table of Contents

Introduction

The modeling and analysis of protein structures and interactions has developed into a major field of research, diverse, difficult, and of exceptional scientific and practical relevance.

Those factors – diversity, difficulty, and relevance – were the motivation for a non-mathematical review (Fraga and Parker 1994), prepared for the *3rd* International Congress on Amino Acids, Peptides, and Analogues, held in Vienna, August 23-27, 1993 [*Amino Acids 5*, 103-216 (1993)]. The interest of many colleagues for that review prompted us to expand it into the present work, maintaining the same objective: to acquaint the reader with the perils and rewards and emphasize the need for a collaborative interaction between experimental and theoretical researchers.

The mixed feelings among researchers were highlighted in that review through a selected sampling of opinions, ranging from an early optimism to a realistic appreciation of the many problems to be faced (Bradley 1970, Wilson and Klausner 1984, King 1989, Wilson and Doniach 1989, Holm and Sander 1992, Ngo and Marks 1992, Honig *et al.* 1993, Eisen *et al.* 1994, Gronbech-Jensen and Doniach 1994, Kolinski and Skolnick 1994, van Gelder *et al.* 1994), with the conclusion that information from as many sources as possible should be used (Thornton 1988), with close collaboration between experimental and theoretical/computational researchers (Daggett and Levitt 1993).

It is this last point that we would like to explore here in more detail. The practitioner in this field might be, in some very special cases, knowledgeable in the life sciences (from biology and biochemistry through pharmacology and immunology), chemistry, quantum mechanics, and computational techniques. In general, however, it will not be so. On one hand, for example, it may be that a worker specialized in computational chemistry will feel tempted to apply his expertise to problems in the life sciences. Two avenues are possible in such a situation: either to dedicate a considerable amount of time in order to become knowledgeable in that new field or to try and engage in a collaboration with an experimental researcher.

Equally serious are the difficulties to be encountered by an experimental worker, who decides to complement/expand the experimental information with computer simulations. Such a researcher would have to decide whether to try and

become an expert in theoretical methods and computational techniques, interact with a theoretician, or rely on the use of existing software.

The difficulties facing both types of researchers, when working on their own, as well as the communication gap existing between them may be brought into focus with the following example (*not* a recommendation) of a possible strategy for the development of a new drug:

> Preliminary work on the target protein (assuming that its tertiary structure is not known) might be carried out using a *symbolic network*, incorporating a *bi-level non-symbolic network* for the prediction of the secondary structure. Actual calculations might then be started with an initial build-up *from fragments*, through an *inverse folding* procedure, or by means of a *genetic algorithm*, followed by a *brute force energy minimization*, with refinement and prediction of additional quantities by either a *Monte Carlo with minimization*, *simulated annealing*, or *molecular dynamics* (in an *essential subspace*) procedure, using an appropriate *potential energy function (AMBER/CHARMm/ECEPP/GROMOS/MM3/?)*, with a *generalized force shifted potential truncation* method for the *non-bonded interactions*. The interaction of prospective *ligands* with the protein could then be studied first within the framework of only the *electrostatic interactions* (from the solutions of the corresponding *linearized Poisson-Boltzmann* equation) and then refined through a *docking* procedure, either making use (with care) of *shape/chemical complementarities* or in a complete treatment (perhaps with *atomic mass weighting*) in either an *explicit* or *non-explicit-solvent* approach. The prediction of *agonists/antagonists/inhibitors* would then be completed with an appropriate *3D-QSAR* methodology (again with care) and the final determination and refinement of the corresponding *peptidomimetics* (if such were the purpose of the study) through quantum-chemical calculations, using the appropriate software (*GAMESS/GAUSSIAN 92/HONDO/MELDF/MOLCAS/?*). These last steps in the study would benefit, naturally, from the information obtained from *combinatorial libraries*.

It is our hope that the present work will be of help in bridging the communication gap between experimental and theoretical researchers, thus encouraging fruitful collaborations. The coverage is wide enough and with sufficient detail to provide a fair picture, with a selected sampling of the literature. Its limitations, which should be evident to the reader, stem from the staggering amount of material appearing in an endless flow, such that each chapter in this work could be easily expanded into an independent monograph.

Lest the reader is given a false sense of optimism, the difficulties and deficiencies have been pointed out again and again, but we would like to stress that there are also success stories (Boyd 1990, Gund 1994), which add further interest to this very exciting field.

Protein Folding

This work would be incomplete without a summary of the basic concepts regarding peptides and proteins to acquaint the reader, who ventures into this field for the first time, with the terminology in use.

The building blocks for peptides and proteins are the *L*- and *D*-amino acids, whether natural or non-natural. The chemical formulas of the twenty natural *L*-amino acids are given in the following chapter and their Cartesian coordinates are presented in Appendix 2.

The peptide chain, formed by amide (peptide) bonds between successive residues is characterized by the appropriate torsion angles (ϕ, ψ, ω, χ), mentioned repeatedly throughout the text, and their definitions may be found in the following chapter.

Designations such as primary, secondary, tertiary, and quaternary structure; α-helices, β-sheets, β-turns, γ-turns; domains; structural motifs; ... are also defined in the following chapter, with graphic representations where appropriate.

In any case, the reader might wish to consult the well-known work of Schulz and Schirmer (1979), Richardson (1981), Ghelis and Yon (1982), Creighton (1983), Chothia (1984), Jaenicke (1987), Valencia Herrera *et al.* (1988), Chothia (1990), and Chothia and Finkelstein (1990), describing in detail the properties of amino acids, the chemical and structural properties of proteins, the energetics of protein conformations, folding patterns, protein biosynthesis, origins and evolution of proteins, ...

2 Amino Acids, Peptides, and Proteins

2.1 Introduction

Proteins play a central role in most biological processes, interacting with *DNA*, *RNA*, other proteins, carbohydrates, lipids, and other organic and inorganic molecules. Proteins transmit chemical and physical signals between molecules in the cell, act as receptors on the cell surface, control the activity of other proteins as well as of *DNA*, transport oxygen, lipids and metals in the blood, act as storage proteins, and control the flow of ions and other molecules across the cell membrane and participate in the transfer of electrons in photosynthesis. There are many proteins that play a major protective role in the immune system, and others that act as important structural and functional components in the cell.

Most of these proteins are the end product of the nucleic acid genetic code. The universality of the genetic code in bacteria and eukaryotes suggests that this relationship between proteins and nucleic acids was established early in evolution. It is this relationship that has led to a long standing debate on the origins of proteins, nucleic acids, and life on earth. Explanations for the origins of amino acids include the prebiotic formation of amino acids by electrical discharge in mixtures of methane, ammonia, hydrogen, and water (Miller 1953), and the formation of peptide bonds by cyanamide (Oro 1963) and reductive acetylation of amino acids associated with pyrite formation (Keller *et al.* 1994); it has also been proposed that adenine was formed from aqueous solutions of ammonia and cyanide (Oro 1960). It has been particularly difficult, however, to explain the formation of the sugar portion of *DNA* or *RNA* and their subsequent polymerization, having been proposed that early sugars were made from glycoaldehyde phosphates (Eschenmoser and Loewenthal 1992) to give phosphorylated ribonucleotides. The appearance of membranes to contain *RNA* or *DNA* is proposed as an important evolutionary step which provided ways to contain and improve the genetic information (Ourisson and Nakatani 1994).

The first step in protein expression is transcription of the *DNA* sequence to the messenger *RNA* (*mRNA*), which requires several regulatory proteins and *RNA*

polymerase. In some cases the *mRNA* is then modified to remove non-protein coding nucleotide sequences (*introns*). Mature *mRNA* is then translated to the amino acid code by interaction with transfer *RNA (tRNA)*, which carry the amino acids that are incorporated into the growing protein chain, and with a complex of ribosomes. Ribosomes are complexes of several proteins and ribosomal *RNA (rRNA)* (Lewin 1990). Proteins are synthesized from the *N*- to the *C*-terminus, which corresponds to the sequence of *DNA* read from the 5' to the 3' end.

It is thought that double stranded *DNA* conserves the genetic code fidelity. It is interesting to note that the codes (Table 2.1) with the most stable nucleoside base pair, *G-C*, code for the single amino acid residues glycine, alanine, proline and arginine. This code is universal but there are some variations in mycoplasma, protozoa, and especially mitochondria. While one strand of the *DNA* code, called the *sense* strand, is used almost exclusively to generate the coded protein, little is known about the purpose of the *antisense* strand of *DNA*. An interesting proposal is that structural information is retained in both sense and anti-sense strands of *DNA* (Table 2.2) (Zull and Smith 1990). It is thought that the protein coding portion of genes is only 3% of the total *DNA* in the 4,000 genes that have been identified of the assumed 50,000-100,000 total human genes. Several points of view exist on whether exons, which encode protein sequences, and introns, which are non-protein coding *DNA* segments inserted between exons, provide evidence to the evolution of protein sequences (Stoltzfus *et al.* 1994). The sequences of proteins have been used to determine evolutionary ancestors and to provide information on their structural and functional properties (Doolittle 1992). A database of 551 ancient conserved regions in proteins (Green, 1994) has been proposed.

Today it is generally thought that *RNA*, which plays an intermediate role in protein synthesis, carried the original genetic information for protein synthesis (Eigen *et al.* 1981, Orgel 1994, Schimmel and Henderson 1994). It is proposed that, because of the stability of the *G-C* base pair, the early triplet codes may have been *GGC, GCC, GAC, GUC* or *GGG, GCG, GAG, GUG*. This means that the original amino acids probably were glycine, alanine, aspartic acid, glutamic acid, and valine, but it might as well be that the evolution of the genetic code is directly related to the precursor-product relationship in the biosynthetic pathway of amino acids where all amino acids originated from Ala, Asp, Glu, Gly, Phe and Val (Wong 1975). Recent information has shown that *RNA* can replicate in the absence of proteins (Kruger *et al.* 1982, Guerrier-Takada *et al.* 1983, Bartel and Szostak 1993) and that it is the *RNA* in ribosomes that catalyzes peptide bond formation (Noller 1993).

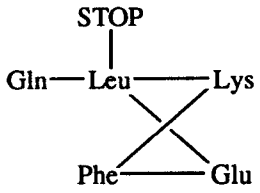
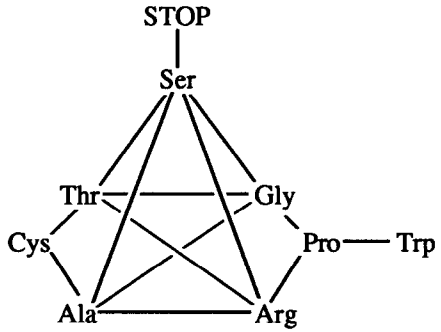
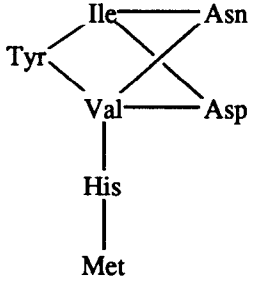
In another question related to the origins of proteins, it is not known how the two types of aminoacyl-*tRNA* synthetases, that specify which amino acid is coupled to *tRNA*, coevolved with protein synthesis and the genetic code (Steitz 1991, Moras 1992, Delarue 1995). Type II synthetases may have evolved first, since they are

Table 2.1. Genetic code for *mRNA* translation to amino acid residue.[a]

	G	C	U	A	
G	Gly	Ala	Val	Glu	G
G	Gly	Ala	Val	Asp	C
G	Gly	Ala	Val	Asp	U
G	Gly	Ala	Val	Glu	A
C	Arg	Pro	Leu	Gln	G
C	Arg	Pro	Leu	His	C
C	Arg	Pro	Leu	His	U
C	Arg	Pro	Leu	Gln	A
U	Trp	Ser	Leu	STOP	G
U	Cys	Ser	Phe	Tyr	C
U	Cys	Ser	Phe	Tyr	U
U	STOP	Ser	Leu	STOP	A
A	Arg	Thr	Met	Lys	G
A	Ser	Thr	Ile	Asn	C
A	Ser	Thr	Ile	Asn	U
A	Arg	Thr	Ile	Lys	A

[a]*G*, *C*, *U*, and *A* stand for guanine, cytosine, uracil, and adenine, respectively. The first column (*G*, *C*, *U*, *A*) represents the nucleoside code for the first letter in the *mRNA* nucleoside code, the top row (*G*, *C*, *U*, *A*) is the second letter code and the last column (*G*, *C*, *U*, *A*) is the third letter code. Note that the strong hydrogen bonded pairs of *G* and *C* for *mRNA* to *tRNA* (*GGX*, *GCX*, *CGX* and *CCX*) code for the single amino acids Gly, Ala, Arg and Pro. Codes with *G* and *C* as the first and last letters (*GXG*, *GXC*, *CXG*, *CXC*) translate to amino acids which can be considered as the basic elements for a stable peptide or protein, including small residues (Gly, Pro, Ala), charged residues (Arg, Asp, Glu), nonpolar residues (Val, Leu) and residues frequently observed in enzymatic sites (His, Gln, Asp, Glu). Tryptophan and Methionine are unique, since they are the only amino acids coded by single triplet *mRNA* codes. The codes *UAA*, *UAG* and *UGA* code for protein chain termination.

Table 2.2. Sense-antisense exchange for codons.[a]

Class	Anticodon Exchange
A	
B	
C	

[a]For example, in class *A*, a sense code for Leu (*CUG, CUC, CUU, CUA*, see Table 2.1) would correspond to codes in the antisense strand for Gln (*GAC*, translated as *CAG* = Gln), Glu (*GAG*, translated as *GAG* = Glu), Lys (*GAA*, translated as *AAG* = Lys) and STOP (*GAU*, translated as *UAG* = STOP). Classes *A*, *B*, and *C* group the α-helical, small and polar (except Arg and Trp), and β-sheet residues, respectively.