Current Topics in Biochemistry

1973

EDITED BY

C. B. Anfinsen

Alan N. Schechter

ACADEMIC PRESS, INC.

A Subsidiary of Harcourt Brace Jovanovich, Publishers

Current Topics in Biochemistry 1973

Edited by
C. B. ANFINSEN
ALAN N. SCHECHTER

National Institutes of Health
Bethesda, Maryland

ACADEMIC PRESS, INC. 1974 New York and London

ACADEMIC PRESS, INC.
111 Fifth Avenue, New York, New York 10003

United Kingdom Edition published by
ACADEMIC PRESS, INC. (LONDON) LTD.
24/28 Oval Road, London NW1

LIBRARY OF CONGRESS CATALOG CARD NUMBER: 72-77222

PRINTED IN THE UNITED STATES OF AMERICA

CONTRIBUTORS

Henryk Eisenberg
Fogarty International Center, National Institutes of Health, Bethesda, Maryland and Department of Polymer Research, The Weizmann Institute of Science, Rehovot, Israel

Donald S. Fredrickson
Molecular Disease Branch, National Heart and Lung Institute, National Institutes of Health, Bethesda, Maryland 20014

David H. Sachs
Immunology Branch, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20014

Harold A. Scheraga
Department of Chemistry, Cornell University, Ithaca, New York 14850

Robert T. Simpson
Section on Developmental Biochemistry, National Institute of Arthritis, Metabolism, and Digestive Diseases, National Institutes of Health, Bethesda, Maryland 20014

E. Brad Thompson
Laboratory of Biochemistry, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20014

Bernhard Witkop
Laboratory of Chemistry, National Institute of Arthritis, Metabolism, and Digestive Diseases, National Institutes of Health, Bethesda, Maryland 20014

PREFACE

This volume is based on a series of lectures held at the National Institutes of Health. This group of lectures is the most recent in a program, which was originated in the mid-1960s, to review various research fields for the scientific community at the Institutes. The topics for these series were chosen to emphasize and summarize active fields of general interest for a diverse audience of scientists. The speakers were encouraged to present an overview of their fields rather than a detailed discussion of current research problems.

The substantial attendance at these lectures, and the response to the two series which have been published in this format, have reinforced our belief that such 'state-of-the-art" reviews are useful to a large number of research workers.

As in the previous published series, this collection covers a wide range of topics, including studies of pure proteins, gene expression in eukaryotes, and metabolic diseases. The unifying theme is the use of chemical methods in studying biological problems.

We again thank Mrs. Anne Ettinger and Mrs. Dorothy Stewart of the Laboratory of Chemical Biology, NIAMDD, for their skilled help in the assembly of the book and we thank the staff of Academic Press for its cooperation.

C. B. Anfinsen
A. N. Schechter

CONTENTS

PREDICTION OF PROTEIN CONFORMATION

Harold A. Scheraga

Department of Chemistry
Cornell University
Ithaca, New York 14850

I. Historical Introduction

In this lecture, I will present a summary of the present status of the problem of predicting the conformation of a protein from a knowledge of its amino acid sequence, and will also mention briefly some preliminary results on the calculation of the preferred conformations of enzyme-substrate complexes.

It is of interest to begin with some historical perspective of the problem. Almost 25 years ago, when Sanger and his collaborators deduced the first amino acid sequence of a protein, insulin (58),--at a time which preceded the determination of the first crystal structure of a protein, myoglobin (32) and hemoglobin (47),--we began to try to determine the structure of a protein in solution by chemical and physico-chemical methods. Our approach was to find the location of many local pair-interactions which would act as constraints on the folding of the polypeptide chain. The covalent structure already provided a knowlege of a few interactions, *viz.*, the specific half-cystine residues involved in disulfide bonds, and we attempted to deduce the location of non-covalent interactions between specific residues. Having available the amino acid sequence of insulin (58), we began with this protein (60, p. 241). However, for a variety of reasons, primarily its insolubility in the neutral pH region, insulin proved to be a difficult protein with which to work. Therefore, with the knowledge that Stein and

Moore and their collaborators (29,66-68) at the Rockefeller Institute and Anfinsen (54) and others (21) at the National Institutes of Health were working on the amino acid sequence of bovine pancreatic ribonuclease, a protein (and, incidentally, an enzyme) with physical properties much more compatible with our experimental approach, our efforts were turned toward ribonuclease (60, p. 270), whose covalent structure is shown in Fig. 1, and also, for similar reasons, to hen egg white lysozyme (60, p. 254).

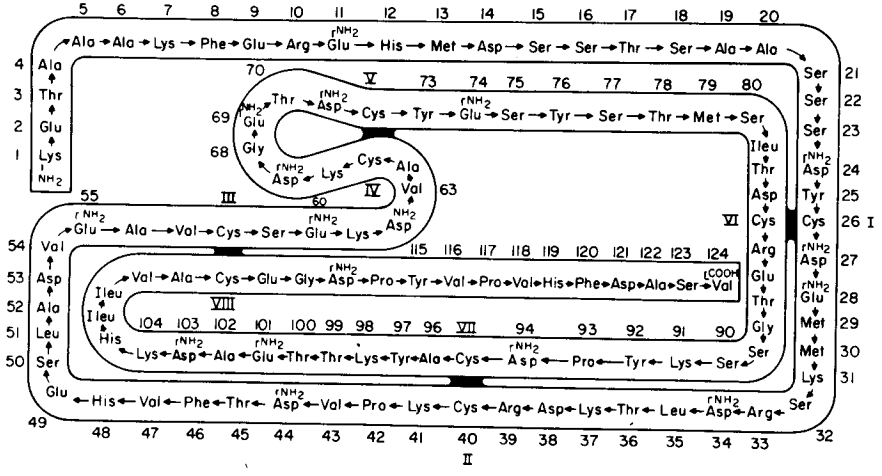

Fig. 1. Amino acid sequence of bovine pancreatic ribonuclease A (21,29,54,66-68).

Over a period of about 10 years, during which most of our efforts were devoted to ribonuclease, we acquired information about the proximities of several residues (61), in addition to the known location (68) of the four bridges between cysteine residues. In particular, three of the six tyrosyl residues were paired specifically with three of the eleven carboxyl groups (all aspartyl residues). Considering that there are 3300 ways to pair 3 of 6 tyrosyls with 3 of 11 carboxyls (61), this specific pairing [which is consistent with

the subsequently-determined X-ray structure of ribonuclease (31,78)] represented the fruitful results of a long series of chemical and physico-chemical studies (61). In the same period, it had been demonstrated that His-12 and His-119 were near each other (5,22,26) and close to Lys-41 (25,28), all three constituting part of the active site. Having available these constraints (4 disulfide bonds and 3 tyrosyl-aspartyl interactions, and the proximity of His-12, His-119, Lys-41), we began to consider how these might be used to determine the three-dimensional structure of the whole protein. At this point, an assumption was introduced, *viz.*, that the native conformation would be the thermodynamically most stable one. Thus, one would make use of the above constraints and try to find the conformation of lowest free energy. The validity of this assumption gained considerable support from the experiment of Anfinsen (4), who showed that (upon reoxidation) reduced ribonuclease (with its 4 cystines converted to 8 cysteine residues) could fold spontaneously to yield the native structure with formation of the correct disulfide bonds. With this background, we began (44) to develop the methods to generate an arbitrary conformation of a protein and compute its energy, and ultimately its free energy, subject to any constraints such as those discussed above, so that one could select out of the enormous number of conformations accessible to the polypeptide chain the one which corresponds to the lowest free energy.

II. Nature of the Problem

To state the problem in another way, one can imagine having a ball and stick (non-space-filling) model of a protein like ribonuclease, *i.e.*, of a connected sequence of amino acid residues (with the disulfide bonds in their proper places). Since the lengths of the sticks are fixed and the holes are drilled in the balls in specific places, the bond lengths and bond

angles are fixed[1] at values set by the manufacturer. Therefore, the only degrees of freedom (in order to change the conformation) are the dihedral angles for rotation about single bonds (with the amide groups maintained in the planar trans conformation[1]). Remembering that a protein of the size of ribonuclease has about 500 single bonds in its backbone and side chains, it is easily seen that, by such rotations, it is possible to generate millions and millions of conformations, only one narrow range of which corresponds to the native protein. In computational language, the dihedral angles are the independent variables for generating any arbitrary conformation, and the energy of each conformation is computed in a search for the conformation of lowest energy (56,59,62); various entropy contributions are included to obtain the free energy (62). It is now possible to minimize the total interaction energy, including *all* pairwise interatomic interactions, with respect to the dihedral angles for a protein of the size of ribonuclease in a reasonable amount of computer time (75). While such computations have been, and are being, carried out, I want to concentrate in this lecture on the as-yet-unsolved problems which must be surmounted before we can predict the three-dimensional structure of a protein solely from a knowledge of its covalent structure.

III. Empirical Energy Functions

Without getting involved in mathematical details, I simply want to say that we have available empirical

---

[1]In a computation, one can vary the bond lengths, bond angles, and planarity of the amide group by introducing appropriate force constants (45,77). While we have allowed for such degrees of freedom, most of our computations have been carried out with fixed bond lengths and bond angles and planar trans amide groups (*i.e.*, rigid geometry), selected separately for *each* type of amino acid.

energy functions, based on pair interatomic interactions, for carrying out such computations, and also procedures for including the effect of hydration. In the absence of solvent, the parameters of the empirical energy functions have been refined (42) by computing the lattice constants, intermolecular binding energies and some intermolecular force constants of a large number of crystals of small molecules, such as, hydrocarbons, carboxyl acids, amides, etc. These functions have also been tested on a variety of model systems. For example, we have computed the preferred conformations of the N-acetyl N'-methyl amides of the twenty naturally occurring amino acids (38), including not only the energy but also the librational entropy in the calculations. The computed relative amounts of the two dominant species, a five-membered and an equatorial seven-membered hydrogen-bonded ring, respectively, agree in general with values deduced from infrared and nuclear magnetic resonance measurements on these compounds in nonpolar solvents[2] (38). As another example, the correct helix sense of a large number of α-helical homopolymers has been computed with these parameters (43). Thus, while there is always room for improvement, it appears that we have available a reasonably reliable set of energy functions for carrying out computations on polypeptides and proteins in the absence of water. While a procedure is available (16) for including the role of the solvent, and has been applied in a number of computations [*e.g.*, the formation of a hairpin turn in a long α-helical section of poly-(L-alanine) in water (65)], we are devoting considerable

[2]Recently, by considering the theoretical results (38) and related experimental data, it has been shown (10) that a third conformation (with no internal hydrogen bond, and designated as a γ conformation; see Fig. 2) can exist to a large extent for amino acids other than glycine. Heretofore (38), the γ and equatorial seven-membered ring conformations have been considered to be the same.

attention at present toward the improvement of the treatment of the solvation of a polypeptide chain.

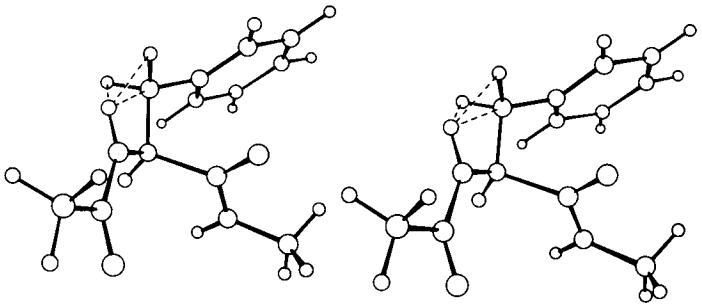

Fig. 2. ORTEP stereo diagram of the dipeptide N-acetyl N'-methyl-L-phenylalanine amide in the γ conformation (ϕ =-60°, Ψ = 140°, χ_1 = 180°, χ_2 = 80°) (10). The dashed lines indicate nonbonded interactions which are thought to affect the N-H stretching frequency.

IV. Multiple-Minima Problem

To turn to the as-yet-unsolved problems, let us recall that a polypeptide of 100 residues has about 500 independent degrees of freedom[1] counting dihedral angles of rotation about single bonds: two per residue in the backbone and an average of about three per residue in the side chain. Thus, the energy surface is a 500-dimensional one and is very complex. To illustrate the difficulty that arises from this complexity, let us pretend, for the sake of drawing a sketch in two dimensions, that the energy, E, is a function of only one variable, q. This dependence might appear like the curve shown in Fig. 3. From this diagram it becomes obvious that conventional minimization procedures will lead to the minimum in the *same* potential energy well as the conformation from which the computation was started. What is required is a procedure to surmount intervening potential barriers (of course, in the 500-dimensional space) in order to reach the *global* minimum, the one of lowest energy or, to reach

the minimum corresponding to the native protein, if it is not the global minimum. The rest of my lecture will be devoted primarily to a description of our efforts to overcome this as-yet-unresolved difficulty, and of the information we have gained during the course of this work about the factors which determine the preferred conformation of a polypeptide chain.

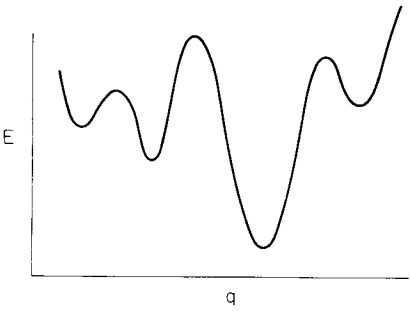

Fig. 3. Schematic two-dimensional representation of the energy as a function of conformation.

Our initial efforts were devoted to the development of mathematical procedures (11-13,17,18) to attain the global minimum. These procedures have proven to be successful for small oligopeptides, but would consume too much computer time for a larger structure. The largest structure to which such procedures have been applied is deca(L-alanine) in water (17). Starting with an α-helical conformation (a local minimum), this decapeptide was carried through 14 successive lower minima (with a net decrease in energy of 38 kcal/mole, without yet reaching the global minimum), the conformation at each successive minimum departing more and more from the initial α-helical structure. It is our view, at the present time, that suitable algorithms are not yet available to provide a mathematical solution to this problem. Therefore, we have had to resort to other procedures to obtain a conformation which would lie in the potential energy well containing the minimum corresponding to that of the native protein. If such

a conformation were attained, presently-available minimization procedures would lead to the minimum.

V. Dominance of Short-Range Interactions

For reasons outlined in section VII, one of the approaches for finding alternative methods led to a consideration of the possible dominance of short-range interactions (33). This investigation led to the concept (33) that the conformation of an amino acid residue in a polypeptide or protein is determined in very large measure, though not exclusively, by the short-range interactions between a side chain and the atoms of the backbone of the *same* amino acid residue, and is, again in first approximation, essentially independent of interactions with neighboring side chains or backbone portions of the chain. This view has recently received further support from a statistical analysis of the conformations of amino acid residues in globular proteins by Finkelstein and Ptitsyn (14). Therefore, let us trace the development and application of this concept, in order to see how it may help us overcome the multiple-minima problem.

VI. Definitions

As used here, the term "short-range" refers to an interaction between the side chain of an amino acid residue with its own backbone. The interaction between the atoms of a given residue with those of any other residue, nearby in the chain or more remote along the chain (even though, possibly, nearby in space) is termed "long-range."

VII. The θ-point

The treatment of an ideal homopolymer chain by random-flight statistics leads to the conclusion that some average linear dimension of the chain, *e.g.*, the root-mean-square end-to-end distance $\langle r^2 \rangle^{1/2}$, varies with the square-root of the molecular weight (15).

While long-range and excluded volume effects (not included in the random-flight calculation) tend to increase $\langle \bar{r}^2 \rangle^{1/2}$ beyond its *ideal* value, the choice of an appropriate (poor) solvent (in which polymer-polymer contacts are favored over polymer-solvent contacts) can reduce $\langle \bar{r}^2 \rangle^{1/2}$ to its *ideal* value (15). Under these conditions (*i.e.*, at the θ-point), the polymer-polymer and polymer-solvent interactions compensate the long-range and excluded volume effects, and the ideal value of $\langle \bar{r}^2 \rangle^{1/2}$ which results is determined entirely by short-range interactions (15). Although a protein in aqueous salt solution may not be at the θ-point, the possibility existed that its conformation, while not determined exclusively by short-range interactions, might nevertheless be dominated by them. As I will show here, the dominance of short-range interactions has been demonstrated for the formation of α-helical and non-helical portions of proteins (14,33), and for the formation of β-turns (39), and also extended structures (9).

VIII. Conformational Preferences Within a Single Peptide Unit

To examine the validity of the hypothesis that short-range interactions are dominant, a study was made (33) initially of the role of these interactions in α-helix formation for proteins of known structure. In particular, calculations were carried out to obtain the energy of interaction (in dipeptide units) of individual side chains in lysozyme with side chains that are nearest neighbors along the backbone, as well as with the backbone groups themselves. It was found that, for various initial backbone conformations (*viz.*, the right- and left-handed α-helices, αR and αL, respectively, and the antiparallel pleated sheet structure, β), the conformation of lowest energy after minimization was the same in most cases for a given amino acid residue and was independent of the nature of the next amino acid in the chain. Furthermore, the backbone structures corresponding to the lowest energy (*i.e.*,

α_R, β, or α_L) showed a high degree of correlation with the so-called helix-making or helix-breaking character of a residue, as determined by earlier *empirical* studies on the identification of α-helical regions in proteins (23,24,55,64). In other words, it appears that the short-range interactions within a given peptide unit may be the physical origin of the so-called helical potential of a residue. In addition, since the side chain-side chain interaction does not play a major role in determining conformation in most cases, the cooperativity among residues, which is necessary for the formation of a helical segment, may simply be the additive effect of placing some sequence of helix-making residues in a particular region. This suggested a model for helix formation in which each type of peptide unit in proteins of known amino acid sequence was assigned a designation *h* or *c* (helix-making or helix breaking, respectively), based on a study of the energy surface of the peptide unit. Then, from an examination of the *h* or *c* assignments for lysozyme, myoglobin, α-chymotrypsin and ribonuclease, empirical rules were formulated to distinguish between helical and nonhelical regions. These rules are: (*a*) an α-helical segment will be nucleated when at least four *h* residues in a row appear in the amino acid sequence and (*b*) this helical segment will continue growing toward the C-terminus of the protein until two *c* residues in a row occur, a condition that terminates the helical segment. With these rules, it was possible to predict the helical or nonhelical state of 78% of the residues of the four proteins mentioned above (33).

With the later availability of the X-ray structures of seven proteins, the validity of these rules was examined further (34). It was observed that, if a nonhelical dipeptide ever occurred at the C-terminus of a helical region, it had a low probability of occurring elsewhere in a helical region and as high as a 90% probability of occurring elsewhere in nonhelical regions; *i.e.*, two *c* residues in a row prevent further growth of a helical segment. It was also found that those residues designated as *c*'s tended to predominate at the C-termini of helical segments. These results constitute

an experimental demonstration of the validity of rule (*b*) above. Finkelstein and Ptitsyn (14) also made a statistical analysis of the conformations of amino acid residues in proteins of known structure, and came to similar conclusions, *viz.*, that short-range interactions are dominant, in that single residues can be classified as helix-making or helix-breaking and that side chain-side chain interactions play a minor role in determining the conformational preference of a given amino acid residue.

At this point, it is of interest to consider the factors which determine the conformational preference of a given amino acid residue. The conformational entropy of a residue in the random coil state must be overcome by favorable energetic factors in order for the residue to be helix-making; otherwise, it will be helix-breaking. Glycyl residues, with no side chains, have no favorable energetic factors to enhance helix formation; thus, the entropy of the coil makes glycyl residues helix breaking (19). When a β-CH_2 group is added, the resulting nonbonded interactions tend to favor the α_R conformation (19,33). Thus, alanine is a helix-making residue (19). While all amino acids besides glycine have a β-CH_2 group, they are not all helix-making because of interactions involving groups beyond the β-carbon; *e.g.*, in Asn which is helix-breaking, electrostatic interaction between the polar side-chain group and the polar backbone amide group de-stabilizes the α_R conformation relative to other conformations. In Gln and Glu, the electrostatic effect is weaker because of the greater distance between the backbone amide group and the polar side-chain group (resulting from entropically-favored extended side-chain conformations); hence, the preferred conformation for Gln and Glu is α_R. Recently, an extensive series of conformational energy calculations (including the computation of statistical weights) was carried out for the N-acetyl-N'-methyl amides of all twenty naturally-occurring amino acids (38). From these calculations, it is possible to assess how the various energetic factors contribute to the conformational preferences of each residue. For example,