PROTEIN STRUCTURAL ANALYSIS, FOLDING AND DESIGN

PROTEIN STRUCTURAL ANALYSIS, FOLDING AND DESIGN

Edited by
Masahiro Hatano
Chemical Research Institute of Non-aqueous Solutions
Tohoku University, Sendai 980, Japan

JAPAN SCIENTIFIC SOCIETIES PRESS Tokyo
ELSEVIER Amsterdam-Oxford-New York-Tokyo
1990

Copublished by
JAPAN SCIENTIFIC SOCIETIES PRESS, Tokyo
and
ELSEVIER SCIENCE PUBLISHERS, Amsterdam

exclusive sales right in Japan
JAPAN SCIENTIFIC SOCIETIES PRESS
6-2-10 Hongo, Bunkyo-ku, Tokyo 113

for the U.S.A and Canada
ELSEVIER SCIENCE PUBLISHING COMPANY, INC.
655 Avenue of the Americas, New York, NY 10010

for the rest of the world
ELSEVIER SCIENCE PUBLISHERS
Molenwerf 1, 1014 AG Amsterdam

ISBN 0-444-81323-3

ISBN 4-7622-3620-9 (Japan)

Supported in part by The Ministry of Education, Science and Culture under Grant-in-Aid for Publication of Scientific Research Result.

Printed in Japan

Preface

Our understanding of the structure-function relationship in proteins was revolutionized in the 1980s by the use of site-directed mutagenesis coupled with X-ray crystallography, which reveals the story of protein functions in their atomic resolution. This was still limited, however, to the hydrolytic enzymes such as lysozymes of smaller molecular weight or to well-recognized hemoproteins such as myoglobin and hemoglobin. Our knowledge of protein-functions is also being increased through such computer-aided approaches as graphic design and molecular mechanics, although these approaches too are currently applicable for smaller proteins

From 1986 to 1988 a group of Japanese scientists pursued research on the structural analyses, folding and design of various types of proteins including cytochrome P-450, ferredoxins, hemoglobin, erabutoxin, neurotoxins, tryptophan synthase, thermophile isopropylmaleate dehydrogenase, fatty acid synthetase, ion-channel proteins and abnormal hemoglobins.

Their findings summarized here provide data on the capability and potential of site-directed mutagenesis, mutations in nature, and computer-aided predictions for more complicated proteins. Our insight into protein structure-function relations is now sufficient to enable us to refine the design of proteins, with the aim of eventually creating protein molecules with specific properties. Publication was supported in part a grant from the Ministry of Education, Science and Culture of Japanese Government, whose valuable support is appreciated.

February 1990 Masahiro HATANO

Contents

II. Higher-Order Structure of Protein Systems

III. Mutation and Structural Design of Proteins

I. PROTEIN STRUCTURAL ANALYSIS AND FOLDING

PROTEIN CONFORMATION IN TERMS OF CONFORMATIONAL ENERGY ANALYSIS

H. Wako

School of Social Sciences
Waseda University
Shinjuku-ku, Tokyo 169, Japan

I. INTRODUCTION

Numerous studies on the anatomy and taxonomy of protein conformations, based on their three-dimensional structures obtained by X-ray crystallography, have been carried out (1). In these studies the three-dimensional structures have been analyzed through visual inspection of the stereoscopic figures, molecular models on a graphic display and so on. In other words, these structures have been inspected mainly from a geometrical point of view. As a result, a static picture of protein conformations has been fixed in the minds of some protein researchers.

Recent computer simulation studies of molecular dynamics and the Monte Carlo method have provided dynamical aspects of protein conformations and have revealed that protein structures fluctuate at room temperature much more than expected (2). However, it is not always easy to obtain such dynamical information, because simulation requires a great deal of computation time, even if supercomputers are used. Since it is important to take account of these dynamical aspects in the anatomy and taxonomy of protein conformations, a convenient method for providing this information is required.

For this purpose, the use of normal mode analysis (NMA) in dihedral angle space has been proposed (3-7). In this report we will discuss what kinds of dynamical information can be derived from a given three-dimensional structure by X-ray crystallography, i.e., from the cartesian coordinate data in the Protein Data Bank (PDB) of Brookhaven National Laboratory through NMA.

A. FEDER System

To analyze three-dimensional structures of proteins in terms of conformational energy function we earlier developed the FEDER program (8). This is based on the ECEPP program developed by the Scheraga group (9), but contains functions for rapidly calculating the first and second derivatives of the conformational energy function. With these derivatives, (a) the conformational energy minimization by the Newton method to obtain the minimum point at

which the second derivative matrix is strictly positive definite, (b) the NMA, and (c) an efficient Monte Carlo simulation (10) can be carried out.

In addition, the FEDER system has the following characters (8): (a) independent variables are dihedral angles. In other words, a molecular model with fixed bond lengths and bond angles is considered. The number of dihedral angles required to describe a given protein conformation (i.e., the number of independent variables) is about one eighth the number of atomic coordinates (i.e., three times the number of atoms). (b) All atoms including hydrogen atoms are taken into account. (c) For rapid computation, an algorithm for computing the first and second derivatives and a program code are produced for a supercomputer.

B. Regularization of Data from the Protein Data Bank

The first step in carrying out the NMA is the regularization of the X-ray crystallographical data. Since X-ray data have slightly different bond lengths and bond angles from the standard values used in the FEDER, conformational energy minimization starting from the X-ray data sometimes leads a calculated conformation to a three-dimensional structure slightly deviating from the X-ray conformation. To obtain the energy minimum conformation very similar to the X-ray conformation, we must perform restrained energy minimization by minimizing the following objective function (3).

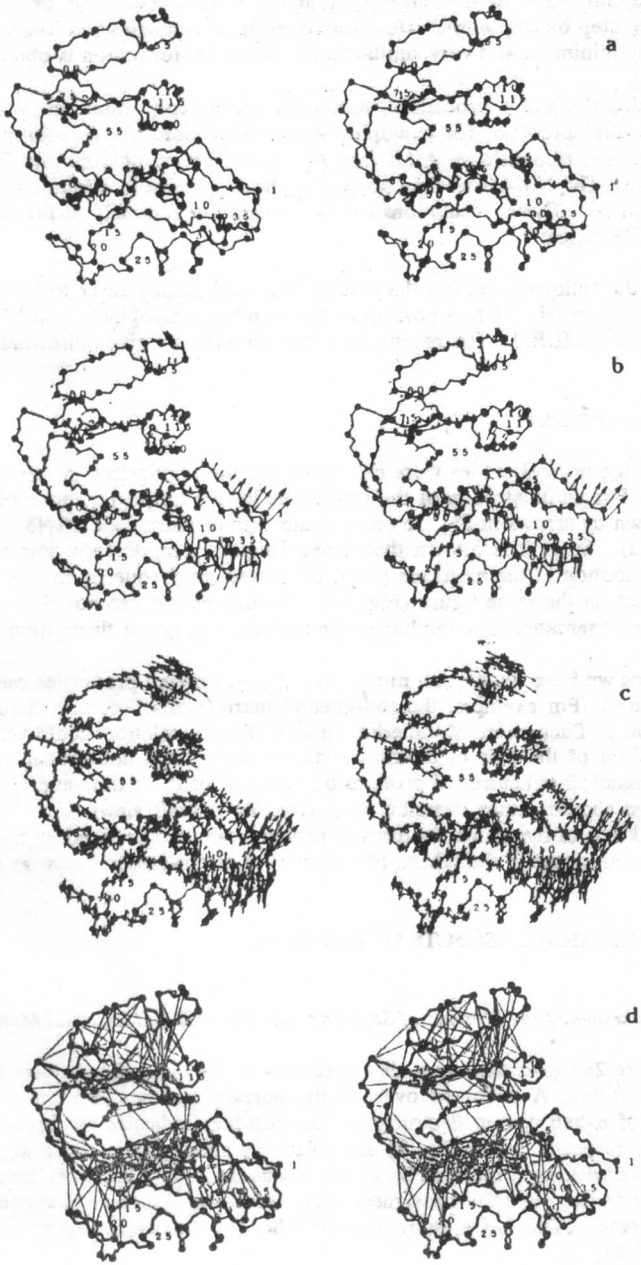
$$F = wp \sum (r_{ij} - r_{ij}^{x})^2 + (1-p)E \qquad (1)$$

where r_{ij} and r_{ij}^{x} are the distances between atoms i and j in the calculated conformation and in the X-ray conformation, respectively; E is conformational energy of the calculated conformation (9); and wp and $(1-p)$ $(0 \leq p \leq 1)$ are the weighting factors for the geometrical constraint and conformational energy terms, respectively (w is in units of kcal/mol/A^2 and p is a dimensionless parameter).

The minimization of the objective function (1) is carried out in several stages as follows: p is initially set at one. After reaching the optimum conformation for this objective function, p is decreased by a certain amount and then

---

Fig. 1. Results for RNase A from normal mode analysis (NMA). In b-d dynamical information obtained from NMA can be added to the structural information obtained directly from the X-ray crystallography in a. (a) Stereoscopic view of minimum energy conformation. (b) Lowest frequency mode. The arrows show the atomic displacement vectors, but their lengths are 7 times longer than those at 300 K. (c) Same as b, but the lowest six modes are also superimposed. (d) The C^{α}-atom pairs with correlation coefficients larger than 0.25 are connected with thin lines (see also Fig. 3a).



the minimization of the objective function is resumed. This procedure is repeated step by step until $p=0$. Consequently, a conformation located at a local energy minimum and very similar to the X-ray conformation is obtained.

Using the above procedure, minimum energy conformations were obtained for several proteins, for example, bovine pancreatic trypsin inhibitor (BPTI), flavodoxin, ribonuclease A (RNase A), lysozyme, myoglobin and Bence-Jones protein. The NMA was then applied to these proteins in minimum energy conformations. These calculations can be carried out routinely at present with the FEDER system.

In the following section the results discussed mainly refer to RNase A in II and BPTI in III. Only a portion of the results on myoglobin and lysozyme are discussed in II.B.3. The results for other proteins are given elsewhere (3,4).

C. Normal Modes

All the normal modes were calculated for the regularized protein conformations. For an illustration of the results the lowest frequency mode of RNase A is shown by arrows in Fig. 1b (coordinate data of entry code 1RN3 in PDB (11) is used). Since it is known that lower frequency modes are dominant in the global motion of the molecule (5-7), the lowest six frequency modes are superimposed on the same figure (Fig. 1c). Though details are not clear in Fig. 1c, such representation can emphasize the regions with larger fluctuation.

Once we have the normal modes, any time-averaged properties can be easily calculated. For example, the root-mean-square (r.m.s.) atomic displacements, the r.m.s. fluctuation of dihedral angles, the correlation coefficients of the movement of the pair of atoms and so on (5-7). It is not enough to describe the dynamical structures of proteins by only one of these time-averaged properties; we should rather describe them from an overall viewpoint. To examine which kinds of dynamical features of proteins can be elucidated by a set of properties calculated in the NMA, two properties are discussed below as examples.

II. DYNAMICAL ASPECTS OF RNASE A

A. Breakdown of the Motions of Segments into Internal and External Motions

Figure 2a shows the r.m.s. displacements of the main-chain atoms (including C^{β}) of RNase A averaged over all the normal modes. The r.m.s. displacements of α-helices and β-structures (the residues indicated by broad and fine stripes, respectively, in Fig. 2) are relatively small, while those of the loops (residues without stripes in Fig. 2) are relatively large (the term "loop" is used to refer to the region not assigned to the secondary structure elements, α-helix or β-strand). This is generally true for other proteins according to our calculations (3,4).

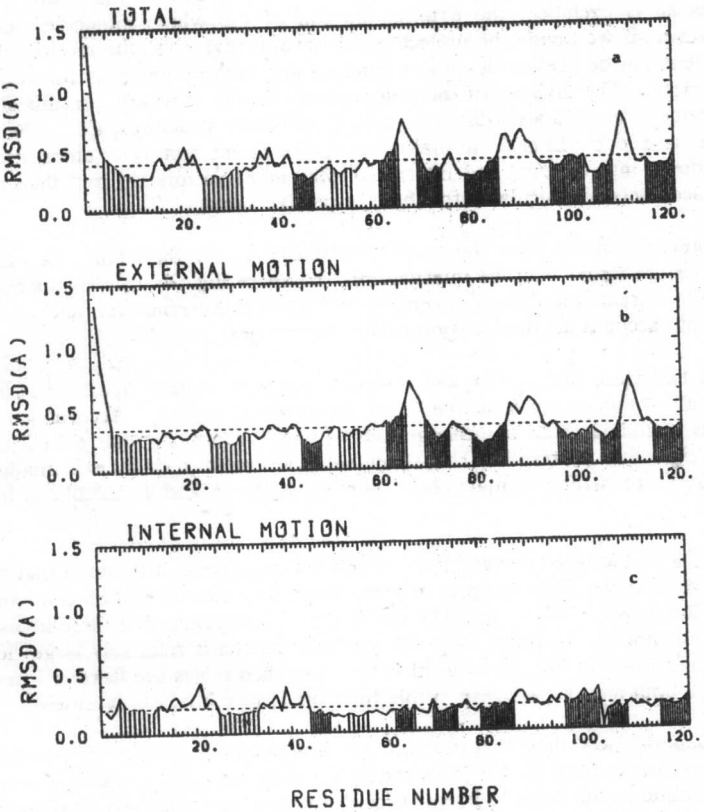


Fig. 2. Root-mean-square atomic displacements for RNase A. Stripes with thick and thin lines indicate the α-helix and β-strand regions, respectively. Other regions are referred to as loops in this report. (a) Values averaged over the main-chain atoms for each residue are plotted against residue number. (b) Displacement for external motion. (c) Displacement for internal motion.

This observation gives rise to the following questions. Does fluctuation of loops come from its internal flexibility, or are the loops rigid with the mobility being due to, for example, the flexibility of the connection region of the loop with the secondary structure element? To answer these questions it is useful to break down the motion of the segment into two types of motions, internal and external. These are defined as follows (7): internal motion is the deformation of a structural element, and external motion is translational and rotational motion, in other words, the fluctuation as a rigid body.

In the NMA the translational and rotational motion of the whole protein molecule is excluded: the external motion of the whole molecule vanishes. However, if we divide the molecule into several segments, the motion of each segment can be broken down into internal and external motions for each normal mode. The division of the polypeptide chain is arbitrary. In this study a molecule is divided according to types of secondary structures, i.e., α-helix, β-strand and loop. In each normal mode displacement vectors of atoms are broken down into external and internal displacement vectors (7) and then r.m.s. displacements are calculated for these respective vectors.

Figure 2b and 2c show the results for RNase A. On the whole, the external motions are larger than the internal ones (compare Fig. 2b, c). This is because the NMA treats the dynamics centered around a local minimum and does not take into account the drastic conformational changes.

As expected, the α-helix and β-strand segments appear rigid, because the internal motion is much smaller than the external motion. According to the results including those of other proteins (3,4), we can say that α-helices are more rigid than β-strands. This statement is also confirmed by examination of the r.m.s. fluctuation of main-chain dihedral angles ϕ and ψ, calculated in the NMA.

As far as the large fluctuations of loops are concerned, it is found that there are several types. For example, in some loops the external motion is dominant (e.g., loops 67-69, 88-97, and 112-117 in Fig. 2), revealing that these loops are rigid but mobile. In other loops the internal motion is relatively large (loops 13-24 and 34-43 in Fig. 2), indicating that they themselves are flexible. In other loops both motions are responsible for their large r.m.s. displacements.

In general, according to our results for several proteins, β-turns connecting the β-strands to form β-structures are rigid, while the loops connecting α-helix and β-strand or the loops in irregular conformations are relatively flexible.

B. Correlation Map of Atomic Movements

1. Correlation Map. Next, we consider dividing a polypeptide chain into several regions such as domains or modules, or detecting core regions defined below from the dynamical point of view rather than from the geometrical one. For this purpose the correlation coefficients γ_{ij} of the atomic motions averaged over all the normal modes were calculated as follows (6):

$$\gamma_{ij} = \langle \delta r_i \cdot \delta r_j \rangle / \langle \delta r_i^2 \rangle^{1/2} \langle \delta r_j^2 \rangle^{1/2} \qquad (2)$$

where δr_i and δr_j are displacement vectors of the atoms i and j from the minimum-energy conformation for a given normal mode and $\langle \cdots \rangle$ means average over all the normal modes and time. If γ_{ij} is positive, it indicates that the pair of atoms i and j has a tendency to move in the same direction as an average over all the normal modes. If γ_{ij} is negative, the two atoms have a

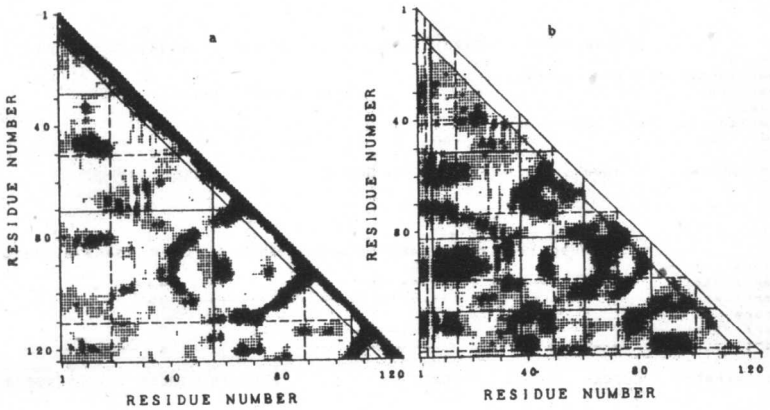


Fig. 3. Positive (a) and negative (b) correlation maps for RNase A. The atom pairs with positive and negative γ_{ij} are marked in a and b, respectively. Atom pairs with $|\gamma_{ij}|$ greater than 0.25 are indicated by squares.

tendency to move in opposite directions, also as an average. However, if the absolute value of the coefficient is nearly zero, there is little correlation between the movements.

In Fig. 3 the results for RNase A are represented in two triangular maps, in which atomic pairs with positive γ_{ij} and negative γ_{ij} are shown separately (referred to as positive and negative correlation maps, respectively). Since the maps for all atoms would be too huge to present here, only the pairs of α-carbon atoms are shown.

In Fig. 1d the positively correlated C^{α} atom pairs with correlation coefficients greater than 0.25 are connected by thin lines. This figure elucidates the clusters of such atomic pairs.

2. *Dynamic Core Region.* A protein molecule is a very complicated system consisting of several thousands of atoms. Therefore, a polypeptide chain must be divided into several regions or characteristic regions identified to derive intelligible information from the three-dimensional structure of the protein.

From a visual inspection of the correlation maps, we are aware of several core regions, which are defined as high density regions of atomic pairs with larger $|\gamma_{ij}|$ values, independent of secondary and tertiary structures.

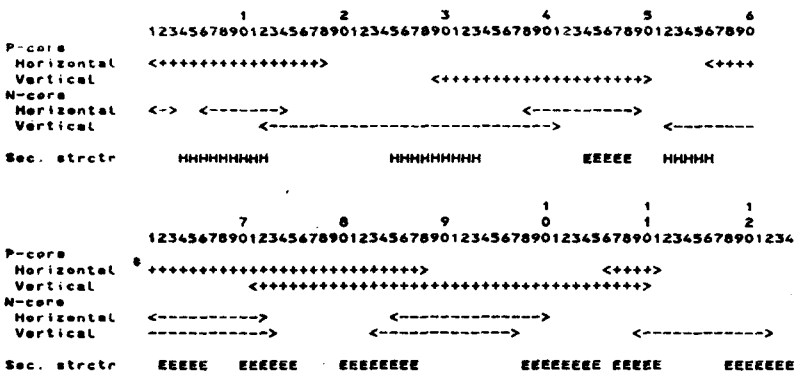


Fig. 4. P-cores and N-cores detected by the horizontal and vertical scanning of Fig. 3a, b. The residues in the α-helix and β-strands are indicated by the letters H and E, respectively.

To objectively define such cores on the maps, the method recently developed by Kikuchi *et al.* (12) for defining domains from the contact maps of proteins is applied (4). By scanning the positive and the negative correlation maps horizontally and vertically with respect to the γ_{ij} values, several core regions are detected and the results are shown in Figs. 3 and 4. These cores detected in the positive and the negative correlation maps are referred to as P-cores and N-cores, respectively.

3. Module. Correspondence of the correlation maps with the modules proposed by M. Go (13) is examined for lysozyme and myoglobin (3,4), and it is found that the negative correlation map has good correspondence with the module. Consequently, the modules can be characterized in terms of the correlation coefficients of the movement of atoms as follows: (a) In a module there exist clusters of atoms in which the movements are positively correlated with each other. This fact corresponds to the compactness of the modules, which is one of the characters originally proposed for a module from a geometrical point of view (13). (b) The atoms in the clusters in a module have negative correlation coefficients of their movements to the atoms in the clusters in other modules. These facts indicate that modules are distinguishable conformational units even from a dynamical point of view.

C. Description of RNase A from a Dynamical Point of View

RNase A is described as follows from a dynamical point of view, according to the above results.

From a visual inspection of the conformation of RNase A (Fig. 1a), two domains linked together near residues 49-51, 76-79, and 102 are well defined. From a dynamical point of view, these domains are also well characterized; for example, they appear as two clusters of positively correlated atom pairs in Fig. 1d. Good correspondence of the domain with the N-cores detected in the negative correlation map (Figs. 3b, 4) reflects the hinge-bending movement of the two domains. This motion is visualized in Fig. 1b,c, in which we observe large displacement vectors at N-terminal and loops 34-40, 90-95 in one domain, and at loops 67-69 and 112-117 in another domain, and very small vectors of this type around the hinge regions.

The P-cores have good correspondence with the secondary structures in general (3,4). For RNase A, the P-core 1-18 contains α-helix 4-12 and the P-cores 56-88 and 71-110 are a part of the β-sheets extending to the two domains.

Fluctuations of the loops are usually large, and are characterized by rigid-like motions for the β-turns 67-69, 88-97, and 112-117 forming the β-sheet, and by the internal flexibility of the loops 13-24 and 34-43 in irregular conformations.

III. ROLE OF DISULFIDE BONDS IN BPTI

A. Perturbation of Protein

The perturbation method is one of the conventional approaches to study very complicated systems such as protein molecules. As a perturbation for proteins, for example, we can consider replacement of some amino acid residues by other such residues, insertion of a small fragment into a certain part of the chain, or deletion of a small fragment from the chain. By studying changes in static and dynamcal conformational features caused by such perturbations we will be able to explore the roles of the replaced, inserted, or deleted amino acid residues in the native conformation.

In this context we studied BPTI. To apply the perturbation to the native BPTI one or more of the three disulfide bonds formed in the native conformation will be unbonded and replaced by cystein residues. By this operation on BPTI we intend to reveal the role of each of the three disulfide bonds in the molecule. Speaking more concretely, if some parts of the chain become more flexible with the unbonding of a certain disulfide bond, such an S-S bond will be seen as having a role in suppressing the fluctuation of these parts in the native conformation.

B. Unbonding of Disulfide Bonds

First of all, the atomic coordinate data obtained from the X-ray crystallography of BPTI (14) were regularized by the FEDER, as described above. The