PROTEINS: Structure, Function, and Genetics 34:369–382 (1999)

Analysis of Domain Motions in Large Proteins

Konrad Hinsen,* Aline Thomas, and Martin J. Field
Laboratoire de Dynamique Moléculaire, Institut de Biologie Structurale— Jean-Pierre Ebel, Grenoble, France

ABSTRACT **We present a new approach for determining dynamical domains in large proteins, either based on a comparison of different experimental structures, or on a simplified normal mode calculation for a single conformation. In a first step, a deformation measure is evaluated for all residues in the protein; a high deformation indicates highly flexible interdomain regions. The sufficiently rigid parts of the protein are then classified into rigid domains and low-deformation interdomain regions on the basis of their global motion. We demonstrate the techniques on three proteins: citrate synthase, which has been the subject of earlier domain analyses, HIV-1 reverse transcriptase, which has a rather complex domain structure, and aspartate transcarbamylase as an example of a very large protein. These examples show that the comparison of conformations and the normal mode analysis lead to essentially the same domain identification, except for cases where the experimental conformations differ by the presence of a large ligand, such as a DNA strand. Normal mode analysis has the advantage of requiring only one experimental structure and of providing a more detailed picture of domain movements, e.g. the splitting of domains into subdomains at higher frequencies. Proteins 1999;34:369–382.**

Key words: protein dynamics; normal mode analysis; domain motions

INTRODUCTION

With more and more structures of large proteins becoming available, there is a growing interest in studying these structures in more detail, both in order to classify repeating structural patterns and to understand the relations between structure and function. A central concept in the study of large protein structures is the "domain," loosely defined as a recognizable substructure within a protein.

There are at least two common definitions for "domains" in proteins. One definition is structural: a domain is defined as a compactly folded part of the protein that is linked to other domains by very few (often just one) structural elements such as a loop or a helix. Various precise definitions and algorithms for finding such structural domains have been published, and a domain database has been constructed by using a consensus procedure based on several domain assignment algorithms.[1] The other definition, and the one we will use here, is dynamical: a domain is a relatively rigid region in a protein that is separated from other domains by more flexible interdomain regions.

In many cases, structural and dynamical domains coincide: since the interactions within a closely packed structural domain are stronger than the interactions between two such domains that are weakly coupled, they are also relatively rigid. However, the two approaches can also lead to very different results, especially in large multi-chain proteins, where dynamical domains can span several chains, whereas structural domains by definition are parts of a single chain. The inverse is also possible: a single structural domain can contain two or more particularly rigid regions that form separate dynamical domains.[2]

The interest in dynamical domains is mainly due to the observation that domain motions are related to the function of many proteins.[2] Several algorithms for detecting dynamical domains in proteins have been published recently,[3–7] and a catalog of domain motion types has been compiled.[8] All of these algorithms require two different conformations as input, and most of them define a rigid domain as a region whose internal geometry is the same (to a precision that is also an input parameter) in both conformations. The only exception is the method by Hayward et al.,[6] which defines rigid domains as regions with constant rotational motion, again up to a specifiable tolerance, and which has also been applied to the analysis of individual normal modes.[5] In practice, all of these methods have some shortcomings that prevent their application to many interesting cases. Specifically, none of them has ever been demonstrated for large proteins and/or proteins with many domains. Moreover, the need for two known conformations limits the application of all these techniques to a small number of proteins, with the exception of the normal mode based technique,[5] which however can only be applied to small proteins.

In this article, we present a new approach that has several advantages. First, it offers a choice of conformation comparison or use of a set of approximate normal modes; these approximate modes can be calculated rapidly and have been shown to reproduce domain motions very well.[9] We will show for three examples that conformation-based and mode-based domain analysis yield essentially the same results. Second, it has no intrinsic limitations. The minimal input requirement is a single conformation for the C_α atoms of the protein backbone, and the size of the proteins that can be treated depends only on available

*Correspondence to: Konrad Hinsen, Centre de Biophysique Moléculaire (CNRS), Rue Charles Sadron, 45071 Orléans Cedex 2, France. E-mail: hinsen@cnrs-orleans.fr
Received 10 August 1998; Accepted 16 October 1998

memory and CPU time. There are no limits on the number of chains or the number of domains as in other algorithms. Moderately large proteins (≈3,000 residues) can be treated on a desktop computer, and even the largest known protein structures (≈8,000 residues) require no more than an hour of CPU time on a well-equipped workstation. Third, our method provides a much more detailed description of low-frequency protein motion than other techniques, decomposing the protein into highly flexible regions, rigid domains, and semi-flexible transition regions. In addition, it provides a measure of deformation at each point in the protein, which permits an assessment of the rigidity of a protein on an absolute scale.

METHODS

Protein Model

Since we are interested in the motion of large domains of large proteins, we use a simplified protein model that consists of only the C_α atoms. Such a model is used frequently in the analysis of experimental structures, and it has been shown recently that it is also sufficient to obtain the low-frequency motions by normal mode analysis.[9]

Finite vs. Infinitesimal Motions

In the analysis of deformations, a distinction must be made between *finite* motions, i.e. the motion of each atom from one definite position to another definite position, and *infinitesimal motions*, which merely describe the relative amounts and directions of all atomic motions for very small changes of the conformation. Finite motions are appropriate for comparisons of experimental structures, whereas velocities and normal mode displacement vectors are infinitesimal motions. The distinction is important since for rotational motions, the directions of atomic displacements in finite motions are different from the corresponding infinitesimal directions. Interpreting finite motions as infinitesimal motions produces an artificial compression, and inversely interpreting infinitesimal motions as finite leads to an apparent inflation. In the following, we will always give both forms of the equations if there is a difference.

Identification of Rigid Regions

The most common method for identifying rigid regions in proteins is the difference-distance plot,[10] which shows the change in all residue pair distances from one conformation to another. Since residue pairs in perfectly rigid regions would show no change at all, small values in the difference-distance plot for all pairs of residues in a well-defined part of the protein indicate a rigid region. However, a difference-distance plot has many shortcomings. It contains no information about the geometric proximity of the residues, making it hard to determine the spatial organization of the rigid regions. Moreover, the differences are not weighted by distance, although a change in a short distance is more significant than the same absolute change in a long distance. Finally, a difference-distance plot provides no quantitative information that could be used to judge the rigidity of a region on an absolute scale. The deformation-plot analysis[11] provides a rigidity measure that is defined per residue, by averaging the elements of a difference-distance matrix along one index using a weight that depends on the residue distance along the chain. However, it is not the distance along the amino acid sequence that is relevant for deformation, but the geometric distance in the structure, and therefore a deformation-plot analysis does not work for proteins with complicated interdomain regions.

A better rigidity measure can be obtained by considering a protein as an elastic object and the difference between two conformations as a deformation. The amount of deformation can be measured by a *deformation energy*. For finite movements, i.e. for the comparison of two conformations, we define the deformation energy on atom I as

$$E_i = \frac{1}{2}\sum_{j=1}^{N} k(\mathbf{R}_{ij}^{(0)})\,[|\mathbf{R}_{ij}^{(0)} + \mathbf{d}_i - \mathbf{d}_j| - |\mathbf{R}_{ij}^{(0)}|]^2, \qquad (1)$$

where $\mathbf{R}_{ij}^{(0)}$ is the distance vector from the position of atom j to the position of atom i in the reference conformation and $\mathbf{d}_i$ is the displacement of atom i between the reference conformation and the comparison conformation. For infinitesimal motions, i.e. normal mode displacement vectors or velocities, we use the limit of infinitesimal atomic displacements $\mathbf{d}_i$,

$$E_i = \frac{1}{2}\sum_{j=1}^{N} k(\mathbf{R}_{ij}^{(0)})\,\frac{|(\mathbf{d}_i - \mathbf{d}_j)\cdot\mathbf{R}_{ij}^{(0)}|^2}{|\mathbf{R}_{ij}^{(0)}|^2}. \qquad (2)$$

These energy expressions correspond to an harmonic force field with a force constant that depends on the interatomic distance in the reference conformation. Since a meaningful definition of deformation energy requires a short-ranged force field with no interactions between potential domains, the force constant $k(\mathbf{R}_{ij}^{(0)})$ must rapidly decrease with increasing interatomic distance. We use the form

$$k(\mathbf{r}) = c\cdot\exp\left(-\frac{|\mathbf{r}|^2}{r_0^2}\right), \qquad (3)$$

where the parameter r_0 describes the range of the force field. The value of the parameter c only influences the global scaling of the deformation energy and can therefore be chosen arbitrarily. The range of the force field must depend on the mechanical model that is used for the protein. Since we will use only the positions of the C_α atoms in our calculations, an appropriate choice of r_0 is 0.7 nm, which has been obtained from a comparison of normal mode calculations with different force fields.[9] Wherever numerical values are given for deformation energies, we have used a value of 47,400 kJ/mol nm^2 for the parameter c, because this value makes the highest vibrational frequencies for a C_α model compatible with the frequencies of similar motions studied with the Amber 94 force field,[12] which was used as a reference in Reference 9.

Obviously other functional forms can be used to measure deformation energy, and a meaningful identification of rigid regions is only possible as far as somewhat different forms lead to essentially the same results. However, the fact that the low-frequency normal modes of a protein, which describe domain motions, hardly depend on force field details indicates that deformation energy measures need not be determined very precisely.[9]

It should be noted that Eqs. 1 and 2 provide absolute measures for deformations that can be compared between different proteins. A suitable threshold defining "rigid regions" can therefore be obtained from a representative sample of proteins and then applied to all other proteins. We have found that a threshold of about 100 kJ/mol leads to a reasonable domain definition for a wide range of proteins. It must be stressed that this energy value was obtained on an arbitrarily chosen energy scale. It should not be compared to energy values obtained in any other way.

A comparison of values for the infinitesimal form, Eq. 2, for normal mode vectors is less evident since the energy depends on the undefined amplitude of the displacement vector. A meaningful comparison between different modes and different proteins can be obtained if the amplitude is defined by

$$\sum_{i=1}^{N} |\mathbf{d}_i|^2 = fN, \tag{4}$$

where N is the number of atoms and f is an arbitrary scaling factor with the dimension of a length squared. We choose the value $f = 1$ nm^2. With this definition, we find that 500 kJ/mol is a suitable threshold for identifying rigid regions. It should be noted that for the moment, we limit ourselves to the analysis of a single normal mode. Since a normal mode calculation provides several modes that are of interest for domain motion analysis, and since it is only the ensemble of low-frequency modes that has a physical relevance, one would like to use multiple modes in the analysis. This will be discussed in the Methods section, "Domain analysis of multiple normal modes."

Normal Mode Analysis

An approximate normal mode analysis method that reproduces the low-frequency domain motions very well with negligible computational cost has been developed recently.[9] The method uses the deformation energy defined in Eq. 2 as an approximate harmonic force field for normal mode analysis. The total deformation energy, defined as

$$E = \sum_{i=1}^{N} E_i \tag{5}$$

with E_i given by Eq. 2, can be written as

$$E = \mathbf{d} \cdot \mathbf{H} \cdot \mathbf{d}, \tag{6}$$

where $\mathbf{d}$ stands for the atomic displacements of all atoms and $\mathbf{H}$ is the matrix of second derivatives of the energy with respect to the displacements. The normal mode analysis consists of the diagonalization of this matrix in mass-weighted coordinates. For large proteins, this diagonalization must be restricted to a suitably chosen subspace of possible motions due to memory and CPU time constraints. We use the Fourier subspace that has been described in Reference 9. With this technique, the normal mode calculation can be performed in a very short time; for example, the normal mode calculation for ATCase (see Results section, "Aspartate transcarbamylase") took 9 minutes on a PentiumPro 200 MHz personal computer.

Comparing Experimental Structures

The deformation analysis described in Methods section, "Identification of rigid regions" can be applied directly to two given experimental structures for the same protein. However, in many cases the results seem incompatible with a simple visual inspection of the two structures. Even if there is a clearly visible rigid-body domain motion, the deformation analysis might show strong local deformations everywhere in the protein. One reason for these deformations are experimental errors in the two structures. Experimental structures do not describe static objects, but are average structures of molecules undergoing thermal motion. Even in principle they can be obtained only to a limited precision, and various statistical and systematic errors in measurement and refinement further increase the "noise" in the final data. Of course, high local deformations do not always indicate experimental errors; they can also be a real feature of the transition between the two structures that are compared. In that case, a description in terms of rigid domains is still justified as long as the large-scale collective motions dominate the transition. Whatever the origin of the local deformations, they must be eliminated before a meaningful deformation analysis can be performed.

To see how the noise can be removed from the input structures, it is useful to assume for the moment that the displacement between the two structures is infinitesimal. It can then be written as a superposition of the eigenvectors $\mathbf{v}_j$ of the matrix $\mathbf{H}$ defined in Eq. 6:

$$\mathbf{d} = \sum_{j=1}^{3N} c_j \mathbf{v}_j, \tag{7}$$

This allows a separation of low-frequency motions (including the domain motions we are interested in) and high-frequency motions, which describe local motions. The noise has high spatial frequency and is thus described by the high-frequency mode vectors, which are not important for domain motion analysis. It can therefore be eliminated by filtering out the high-frequency motions.

From Eq. 6 it follows that the gradient of the total deformation energy is given by

$$\frac{\partial E}{\partial \mathbf{d}} = 2\mathbf{H} \cdot \mathbf{d} = 2 \sum_{j=1}^{3N} \lambda_j c_j \mathbf{v}_j, \tag{8}$$

where λ_j is the eigenvalue of **H** corresponding to the eigenvector $\mathbf{v}_j$. If there are significant high-frequency noise contributions, indicated by large values of the c_j that correspond to large λ_j, they make up for the largest contributions to the gradient. Consequently the high-energy noise can be eliminated by a steepest-descent minimization.

This conclusion also holds for finite displacements: a steepest-descent minimization removes the localized noise first and only then reduces the low-frequency domain motions. Such a minimization is therefore a good method to eliminate noise. The precise amount of minimization is not critical; even if the low-frequency domain motions are somewhat reduced as well, their characteristics remain. The major risk in overminimizing is not the loss of the domains or domain motions, but the loss of differences in flexibility in the interdomain regions. In practice, one usually obtains good result by minimizing until the RMS difference between the two conformations is 0.01 nm smaller than it was initially. Low-resolution structures may require further minimization to show a discernible domain structure.

Rigid-Body Motion

In general, the rigidity analysis described in the Methods section, "Identification of rigid regions" will show relatively rigid regions connected by more flexible parts. There may also be flexible pieces at the exterior of the protein, e.g. flexible loops. The rigid regions essentially define the dynamical domains. Obviously such a definition does not yield precise boundaries for the domains, and neither does it allow a division of the complete protein into rigid domains. This, however, reflects physical reality; a protein does not consist of perfectly rigid regions connected by simple mechanical joints, but rather of relatively rigid regions whose relative motion requires the deformation of the more flexible regions between them.

Although two rigid regions separated by flexible parts must constitute two distinct dynamical domains, the absence of a flexible connection region does not necessarily prove that a rigid region forms a single domain, because there could be sharp transitions from one domain to the next that are not captured by the large-scale deformation energy measure. A complete domain analysis therefore requires a further step, namely the calculation of the rigid-body motion of each rigid region. Only if the rigid-body motion of several rigid regions is approximately equal, can they be considered to form a single dynamical domain. In addition, the calculation of the rigid-body motion provides a more detailed description of the domain motions.

For finite displacements, the motion of a rigid body is most conveniently expressed by a rotation around some reference point (we choose the coordinate origin) plus a translation,

$$\mathbf{d}_i = \mathbf{T} + \mathbf{D} \cdot \mathbf{R}_i, \tag{9}$$

where $\mathbf{d}_I$ is the displacement vector of atom i, $\mathbf{R}_i$ is its position, **T** is the rigid-body translation vector and **D** is an orthogonal rotation matrix, which can be parameterized in several ways, e.g. by a rotation axis (a direction in space) and a rotation angle. Thus, in total there are six independent parameters, corresponding to the six degrees of freedom of a rigid body. Since a rigid region as defined above is not a rigid body in the mechanical sense, i.e. an object without any internal degrees of freedom, any given set of displacement vectors between two conformations will describe a combination of a rigid-body motion and internal deformations. The rigid-body motion can be extracted by a translational and rotational fit, for which we use the quaternion-based method described in Reference 13 due to its numerical convenience.

The most general infinitesimal motion of a rigid body is conveniently described by a different set of six parameters, namely the vectors of infinitesimal rotation $\mathbf{\Phi}$ and translation **T** , defined by

$$\mathbf{d}_i = \mathbf{T} + \mathbf{\Phi} \times \mathbf{R}_i. \tag{10}$$

For a set of displacement vectors that do not describe a pure rigid-body motion, the parameters $\mathbf{\Phi}$ and **T** are obtained by a straightforward linear least-squares fit.

Since a rigid-body motion has six degrees of freedom, it is well-defined only for collections of three or more points. The first step in a domain analysis is therefore the division of the whole system into small units of at least three atoms. We choose a simple geometric division of the protein into small cubes with an edge length of 1.2 nm; such a cube contains on average about 6 C_α atoms. Cubes with less than three C_α atoms are discarded, as are cubes whose average deformation measure exceeds a certain threshold. For each of the remaining ones, the rigid-body motion parameters are calculated; in the case of normal mode analysis, the rigid-body motion is determined for each mode separately. For cubes that belong to the same domain, the parameters should be very similar, whereas there should be clear differences between regions that belong to different domains. The domains can therefore be identified by a cluster analysis of the rigid-body parameters.

The division into cubes has two side effects: first, it is in general not compatible with any symmetry that the system might have, and the final domain decomposition therefore does not respect that symmetry. This is not important in practice, since perfect symmetry is an artifact of crystallization anyway; in real conditions, proteins undergo constant thermal motion and are thus close to, but never exactly in, a symmetric conformation. Second, the elimination of cubes with less than three atoms means that some residues might not be recognized as parts of a domain. However, this is not an important problem for our purposes; we are looking for large regions which have no precise boundaries anyway, and therefore the omission of a few residues has no serious consequences. If desired, the omitted regions could be checked for dynamical conformity

with any of the domains after the domain decomposition has been completed.

Finally, it should be pointed out that the elimination of flexible regions from the domain analysis is not strictly necessary; rigid-body motion parameters are mathematically well-defined for any set of three or more atoms, whatever their internal deformation is. However, it is physically unreasonable to analyze strong deformations in terms of rigid-body motions, and the rigid-body motion parameters of deformed regions typically fall in between well-defined domains, making the subsequent clustering procedure more difficult. Moreover, the initial identification of flexible regions should be seen as an advantage, because it provides a more detailed description of the low-frequency motions.

Domain Analysis

The rigid-body parameters defined in the last section can be considered as points in a parameter space. This parameter space is 6-dimensional for a comparison of experimental structures and 6M-dimensional for a domain analysis based on M normal modes; there is one point per rigid (cubic) region. The points describing a single domain must be close together, whereas points from different domains should be well-separated. In other words, the points form clusters in the parameter space, with each cluster corresponding to one domain. In general, there is not a set of well-defined clusters, but a hierarchy of clusters: points that form a single cluster when seen on a large scale are subdivided in subclusters on a smaller scale. It is also common to find points in relatively large clouds between well-defined clusters; these points describe transition regions between domains that were not eliminated by the rigidity criterion because their overall deformation is small.

Many cluster definitions and clustering algorithms have been proposed and used,[14] but there is no single one that works well for all applications. Moreover, we are not aware of any published algorithm that allows a characterization of relatively rigid transition regions. We have therefore used a simple procedure that has the merit of being easy to understand, and has worked well for a large range of examples.

The simplest form of cluster analysis is inspection by eye of a suitable graphical representation of the data. A useful representation for a cluster analysis in any number of dimensions is a parallel-axis plot[15] (see Fig. 3 for an example). In such a plot, each variable is shown on a separate axis, all of which are drawn parallel to each other. Each point is then represented by a line that joins all the axes. A cluster of points is recognizable as a narrow band of lines that stay close together on all axes. Although a detailed cluster analysis for a large protein by visual inspection is tedious and prone to inconsistencies, a graphical representation is recommended as a "sanity check" on the outcome of automatic clustering algorithms, since all clustering algorithms perform badly for some unusual input data.

Any automatic clustering algorithm requires a definition of the distance or similarity between two points. For infinitesimal motions, a distance measure for the translation and rotation parameters separately is the length of the vector difference divided by the length of the vector sum; the inverse of this quantity describes the similarity. The two similarities can be combined into a total one by using a weighted sum. Empirically, the rotation parameter is usually a better domain identifier than the translation parameter, and therefore we define the similarity of two points i and j by

$$S_{ij} = 3\,\frac{|\mathbf{\Phi}_i + \mathbf{\Phi}_j|}{|\mathbf{\Phi}_i - \mathbf{\Phi}_j|} + \frac{|\mathbf{T}_i + \mathbf{T}_j|}{|\mathbf{T}_i - \mathbf{T}_j|}\,. \tag{11}$$

The same definition is used for finite motions, where we define the rotational parameter $\mathbf{\Phi}$ as the product of rotation angle and normalized rotation axis vector. We do not follow Hayward et al.[5] and eliminate translation altogether from the similarity criterion; in large proteins, relative translation can be the most pronounced difference in the motion of two domains, e.g. in the case of ATCase (see Results section, "Aspartate transcarbamylase").

The largest number in the similarity matrix S_{ij}, whose indices we call $i_{\max}$ and $j_{\max}$, identifies the two points that are closest to each other. They are the obvious candidates for defining the beginning of a cluster. We define the complete cluster by all points k for which $S_{ik} > S_{i_{\max}j_{\max}}/c$, where $c > 1$ is a parameter describing the coarseness of the cluster definitions. The points belonging to the newly-formed cluster are removed from the similarity matrix, and the remaining matrix is used to identify the next cluster. The procedure is terminated when all points have been assigned to a cluster.

This procedure yields the best clusters first, with subsequent clusters being less and less well-defined. The last clusters may not be useful in the definition of domains, describing isolated rigid regions that do not belong to any domain. However, no attempt is made to decide algorithmically which clusters correspond to domains. This decision is best taken upon a visual inspection of the clusters.

The coarseness parameter c defines the level of detail in the cluster analysis. Values close to one will yield many small clusters, whereas larger values result in fewer larger clusters that are made up of several smaller ones. Repeating the cluster analysis with different values of c allows the identification of cluster hierarchies (i.e. large clusters that can be divided into subclusters when finer details of the motion are studied). The variation of the domain structure with changing coarseness parameter is also important to distinguish between stable domains and relative rigid interdomain regions. The former grow slowly with increasing value of c until a saturation is reached; the latter decrease at the same time and may disappear completely, subdivide in nonsystematical ways, or coalesce with neigh-

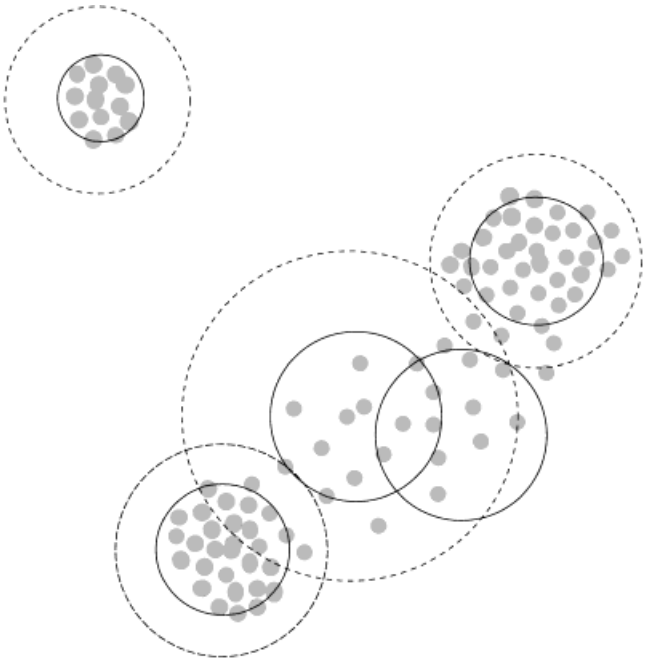

Fig. 1. A schematic illustration of the clustering process. Each grey circle represents the rigid-body motion parameters of one cubic region. The drawn-out circles show the clusters at a small value of the coarseness parameter *c* (discussed in the Methods section, "Domain analysis"), and the dashed circles show the clusters for a higher value of *c*. The isolated group in the upper left corner represents a well-defined dynamical domain, the remaining points can be viewed as two domains with a transition region. A variation of the coarseness parameter *c* affects the transition region much more than the well-defined domains; the clustering as a function of *c* thus provides a way to distinguish between real domains and low-deformation transition regions.

boring regions. This behavior is illustrated schematically in Figure 1.

It should be noted that the cluster analysis procedure described above does not use any geometrical information about the protein. Specifically, it does not define clusters as necessarily contiguous regions in space. If a reasonably complete dynamical description (such as normal modes) is used for finding the domains, no two separated regions should have exactly the same global motion, and this can be used as a verification of the analysis. In the case of a comparison of two configurations, it is possible in principle (but improbable) to find two domains with exactly the same motion, and in that case the coincidence of the global motions is a useful information that should not be eliminated by requiring domains to be contiguous.

Domain Analysis of Multiple Normal Modes

The analysis procedure described in the preceding sections is based on one atomic displacement vector, i.e. the difference between two conformations or a single normal-mode vector. In the case of a normal-mode-based analysis, it makes little sense to restrict oneself to a single mode. It is well known that the precise identity of normal modes depends on force field details that are beyond physical validity. For example, a different treatment of the long-ranged electrostatic interactions will yield different sets of low-frequency normal modes, and no physical argument can be given to decide which treatment is "better." However, the wider subspace of multiple low-frequency modes is unaffected by such details, and in fact can be obtained with drastically simplified force fields and protein models.[9] A domain motion analysis should therefore be based on multiple modes. An added advantage is that multiple modes provide a more detailed description of the dynamics.

The extension of the analysis procedure to multiple mode vectors is straightforward. The deformation energies (Eq. 2) are calculated for each mode, and the rejection criterion for deciding which of the small cubic regions to keep for the domain analysis is applied to each mode as well, i.e. to pass as sufficiently rigid, a region must be below the rigidity threshold in all of the modes. The similarity measure (Eq. 11) is also evaluated for each mode separately, and then a total similarity measure is defined as the sum of the similarity measures of each mode.

The remaining problem is to decide how many modes to use in the domain analysis. Obviously a physically meaningful dynamical description should not depend on any arbitrary choice. It turns out that the number of modes in the analysis does not influence the outcome significantly. Additional modes provide a more detailed picture, e.g. the division of large domains into subdomains. On the other hand, they add more localized motions and hence higher local deformations. The rigidity threshold must therefore be higher, making it more difficult to eliminate flexible regions. As a rule of thumb, a good starting point is to use a rigidity threshold of 400 to 500 kJ/mol and to include all modes whose average deformation energy is below this value.

Characterization of Domain Motions

When the clusters and hence the domains have been determined, the final rigid-body parameters are obtained by another fit for each entire domain. It is then useful to determine the "proper" axis of the motion. Any rigid-body motion can be described as a combination of a rotation around a fixed axis in space and a translation along this axis; this combination is often described as a screw motion around the axis. The direction of this axis is given by the rotation parameter, such that its complete definition requires only the construction of an arbitrary point on it. For infinitesimal motions, such a reference point is given by

$$\mathbf{R}_{\mathrm{ref}} = \frac{\mathbf{\Phi} \times \mathbf{T}}{|\mathbf{\Phi}|^2}. \tag{12}$$

For finite motions as described by Eq. 9, a reference point can be obtained from

$$\mathbf{R}_{\mathrm{ref}} = (\mathbf{D} - \mathbf{1})^{+} \cdot (\mathbf{nn} - \mathbf{1}) \cdot \mathbf{T}, \tag{13}$$

where $\mathbf{n}$ is the normalized direction of rotation and $\mathbf{A}^{+}$ denotes the generalized inverse of the matrix $\mathbf{A}$. In both

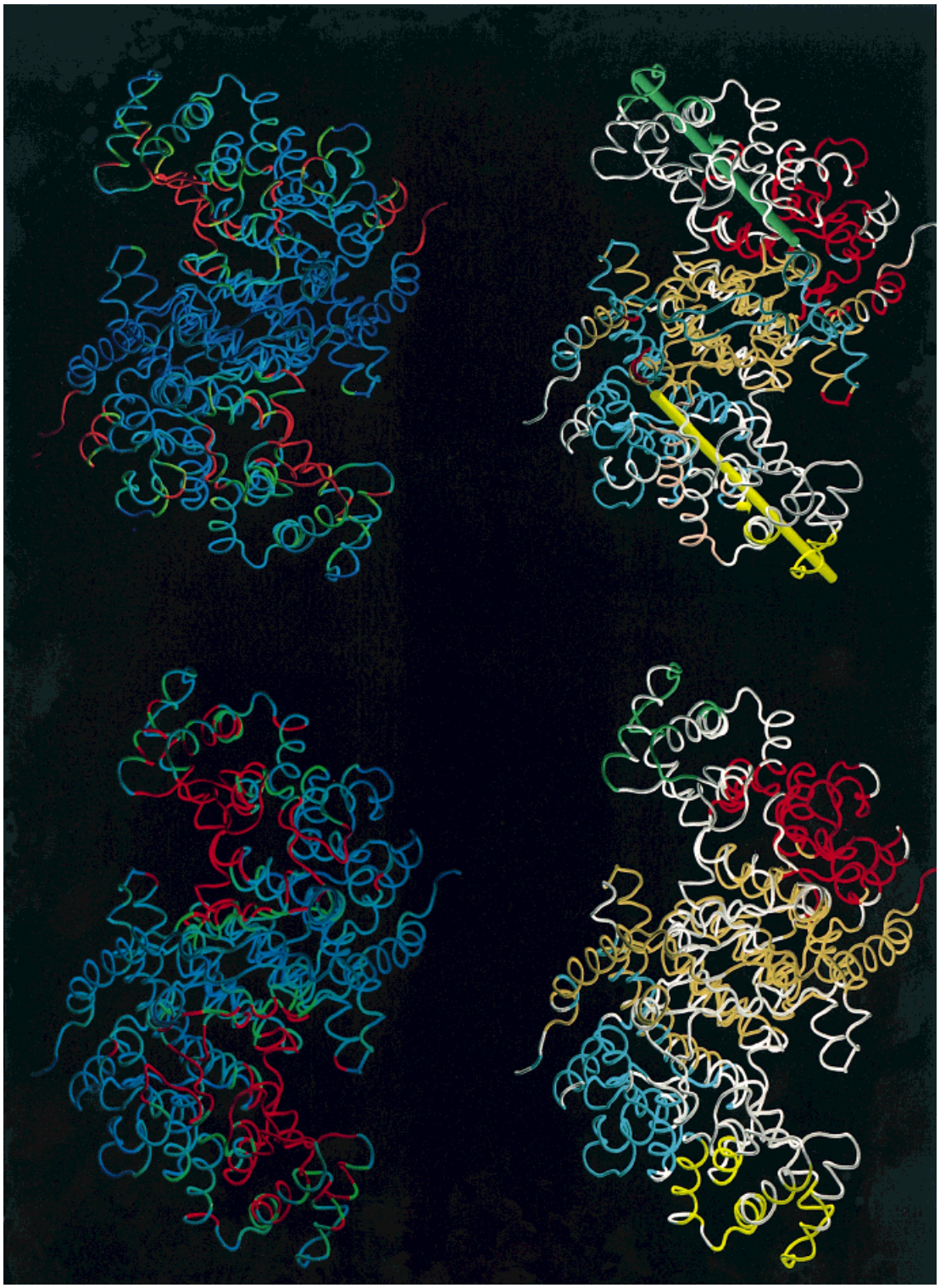

Fig. 2. The deformation energy (left column) and domain decomposition (right column) for citrate synthase, from conformation comparison (top) and normal modes (bottom). Blue regions in the deformation energy pictures are the most rigid ones, red indicates strong deformation. Green corresponds to the rigidity threshold used in the domain analysis. In the domain decompositions, the central orange region and the tips in yellow and green represent stable dynamical domains, the red, cyan, and pink parts are low-deformation interdomain regions. The absence of the molecular symmetry in the domain decompositions is an artifact of the analysis procedure, as explained in the Methods section, ''Rigid-body motion''. The axes shown in the top right picture represent the rigid-body screw motion of the tips relative to the central domain.

cases, the amount of translation along the axis is given by

$$T_{\parallel} = \frac{\mathbf{T} \cdot \mathbf{\Phi}}{|\mathbf{\Phi}|} . \quad (14)$$

The amount of rotation is given by $\phi = |\mathbf{\Phi}|$ for infinitesimal motions and by the rotation angle ϕ for finite motions. The kind of motion can be characterized by the ratio $2\pi T_{\parallel}/\phi$, which is the amount of translation corresponding to a rotation of 2π. A value of zero indicates pure rotation, whereas an infinite value indicates pure translation.

The domain motion description given in this section is also applicable to relative motions between two domains; such a description is often more convenient than a specification of the absolute motion of each domain, especially for proteins with few domains.

Choice of the Reference Structure

The comparison of two given conformations of a protein is at first sight a symmetric problem; interchanging the two conformations could seem to make no difference. This is however not true; it is clear from Eq. 1 that only one of the two conformations is used for evaluating the force constants and hence the weight of each atom pair in the deformation energy. This raises the question if one or the other conformation is preferable as a reference for the domain analysis. Tests on several proteins have shown that both choices yield almost identical domain assignments, but that the domain delimitations are clearer if an open structure is used as reference, i.e. a structure which leaves more freedom for the domains to fluctuate. A more mathematical formulation is that each short-distance atom pair in the reference structure should also be a short-distance pair in the comparison structure, which means that the ideal reference structure is the one with the smallest number of short-distance atom pairs. Yet another way to state this criterion is that the ideal reference structure has the largest possible exposed surface.

The same observation holds for normal mode based domain analysis. Although test calculations show that normal mode calculations on all available conformations of a protein yield similar results, a clearer picture is obtained for conformations with few short-distance atom pairs and a large exposed surface.

RESULTS

In this section, we will present a domain analysis for three proteins: citrate synthase, HIV-1 reverse transcriptase, and ATCase. For each of these proteins, at least two different conformations are available experimentally, and we compare the domain assignment obtained by comparing the conformations to the one obtained from normal mode calculations.

The analyses were made using an interactive analysis program which implements the techniques described in the Methods section. This program, written in a combination of Python and C and making heavy use of the Molecular Modeling Toolkit,[16] is available free of charge for downloading.[17] It can be used on any Unix system running the X window system. The calculations were made using a Hewlett-Packard Vectra VA computer (Pentium-Pro at 200 MHz, 64 MB of RAM) running the Linux operating system; none of them took more than a few minutes.

The images in Figures 2–4 were created using the MMTK toolkit,[16] the visualization program VMD,[18] and the rendering program POV-Ray.

Citrate Synthase

The first example, and with twice 437 residues also the smallest protein we will study, is citrate synthase, which was chosen because it has a relatively simple domain structure, and because dynamical domain assignments have been published before using different techniques.[6,8] Furthermore, a normal mode study[19] has shown that the low-frequency modes compare well with the difference between the crystallographically-observed conformations. Citrate synthase catalyses the formation of citrate in the citric acid cycle and occurs in almost all organisms. It is a symmetric homodimer which is folded almost entirely in an α-helical structure. Each chain comprises two structural domains: a small domain made up of five α-helices, and a large domain consisting of 15 α-helices. Binding of coenzyme A to citrate synthase induces the closure of the cleft between the two domains, where the substrate is located. This motion has been characterized previously as a rigid-body rotation of the small domain relative to the large domain by 18 degrees, with a hinge located close to residues 274 and 275.[20,21] In the classification of domain motion mechanisms by Gerstein et al.,[8] citrate synthase is listed as a typical example of shear motion, as opposed to hinge motion. However, as has been noted before,[6] this classification is based on the structure of the interdomain region, not on the effective relative motion of the domains, which is well described by a hinge model.

Both an open (1CTS) and a closed (2CTS) conformation of citrate synthase are available in the Protein Data Bank;[22] their C_α-based RMS distance is 0.31 nm. Due to the symmetry of the dimer, it would seem sufficient to study the monomer only, and in fact this has been done in an earlier study.[6] However, the interactions between the two monomers are of the same order of magnitude as the inter-residue interactions that cause the tertiary structure. A dynamical domain might therefore contain parts of both chains, which means that the domain analysis should be performed on the dimer. The normal mode calculation has been done for the dimer as well; a Fourier basis of 690 modes (corresponding to a cutoff of 1.17 nm) was used. The first four non-zero modes were used for the domain analysis, using a rigidity threshold of 500 kJ/mol. The rigidity threshold for the comparison of 1CTS and 2CTS was 100 kJ/mol. The cluster coarseness used for Figure 2 was 7 for the normal modes and 8 for the conformations.

Figure 2 shows the result of the deformation (left column) and domain (right column) analysis; the comparison of the two conformations is shown on top and the normal mode results at the bottom. The deformation

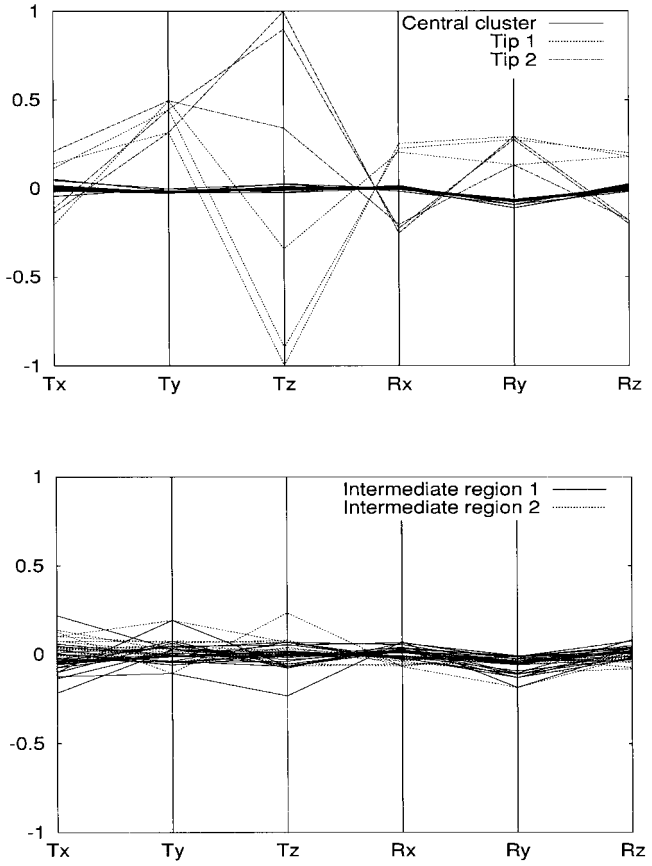

Fig. 3. A parallel-axis plot showing the rigid-body motion parameters for each of the small cubic rigid regions in the conformation comparison of citrate synthase. The six vertical axes represent the six degrees of freedom; the three axes on the left show the translation, and the three axes on the right the rotation. The upper picture shows the well-defined clusters, for which the relative distance between the different cubic regions is small. The lower picture shows the two intermediate transition regions, whose motion is similar to that of the central domain, but shows a much larger spread.

analysis shows a large and very rigid central region, consisting of parts of both chains, and two more flexible "arms" terminating in small more rigid tips. This result is obtained both from a comparison of the two conformations and from a normal modes analysis of the open conformation; the deformation pattern is very similar. The similarity remains in the domain analysis: there are three stable domains (the two tips in green and yellow, and the large central orange region) and two sufficiently rigid connecting regions (red and cyan) which with decreasing values of the domain coarseness show varying subdivisions. The rigid-body motion parameters of the small cubic regions that make up the domains are shown in Figure 3. It clearly shows that the flexible arms have a much wider spread of motion parameters than the well-defined domains. In addition, it shows that the small tips move much more than the connecting regions (the central region is used as the reference coordinate frame and therefore does not move by definition). For the structure comparison, the movement of the tips relative to the central domain is a rotation of 26 degrees around and a translation of 0.07 nm along the axes indicated by the green and yellow cylinders in the picture. The axes pass within 0.25 nm or less from residues 57–58, 64, 378–380, and 383.

It is interesting to compare our domain decomposition to a recently published one[6] that uses a method which is similar in spirit to the one presented here, but different in many important details. Essentially, the global rotation is calculated for short main-chain segments and a cluster analysis yielding many small domains is performed on the rotation vectors. The small domains are then combined into larger ones according to the ratio of interdomain to intradomain motion. The translational motion is not used at all, and there is no initial rigidity criterion to eliminate flexible regions. The procedure was applied to the citrate synthase monomer, using the same input structures that we used in our analysis. The result is a decomposition of the monomer into two domains, a small one that includes the tip described above, but is substantially larger, and a big one, which consists of half of our central domain (the part belonging to the chain under consideration) and the large non-flexible interdomain region. The relative motion of these domains is described as a rotation of 19.2 degrees plus a translation of 0.001 nm with an axis very close to ours. It is not surprising that we find a larger motion, since our reference is just the most rigid central part of the protein.

The main difference between the two analyses is thus the domain decomposition, which is due to both different criteria and different analysis techniques. However, it is clear from Figure 3 that a larger tolerance for the central domain would lead to an inclusion of the intermediate connection regions into this domain, and thus to a domain decomposition essentially equivalent to that given in Reference 6. However, such a decomposition would not describe the existence of a well-defined core of higher rigidity in the center.

This example shows clearly that dynamical domains are not well-defined intrinsic properties of protein dynamics, but rather a useful concept in the description of protein motions. Their interpretation must take into account the form of the description that was used to identify the domains. Our description was designed to provide as much information as possible within the limitations of a decomposition into geometrical regions.

HIV-1 Reverse Transcriptase

Although only moderately larger then citrate synthase, HIV-1 reverse transcriptase (HIV1-RT) has a rather complex domain structure, which makes it an interesting candidate for domain decomposition techniques. HIV1-RT is a polymerase responsible for replicating the single-stranded RNA genome to double-stranded DNA and therefore a potential target for HIV drugs. It is a heterodimer consisting of two differently folded chains (named p51 and p66) whose sequence is for a large part identical.[23,24] From a structural point of view, the chains can be divided into domains that have been named according to their shape resemblance to a right hand; both chains have the domains

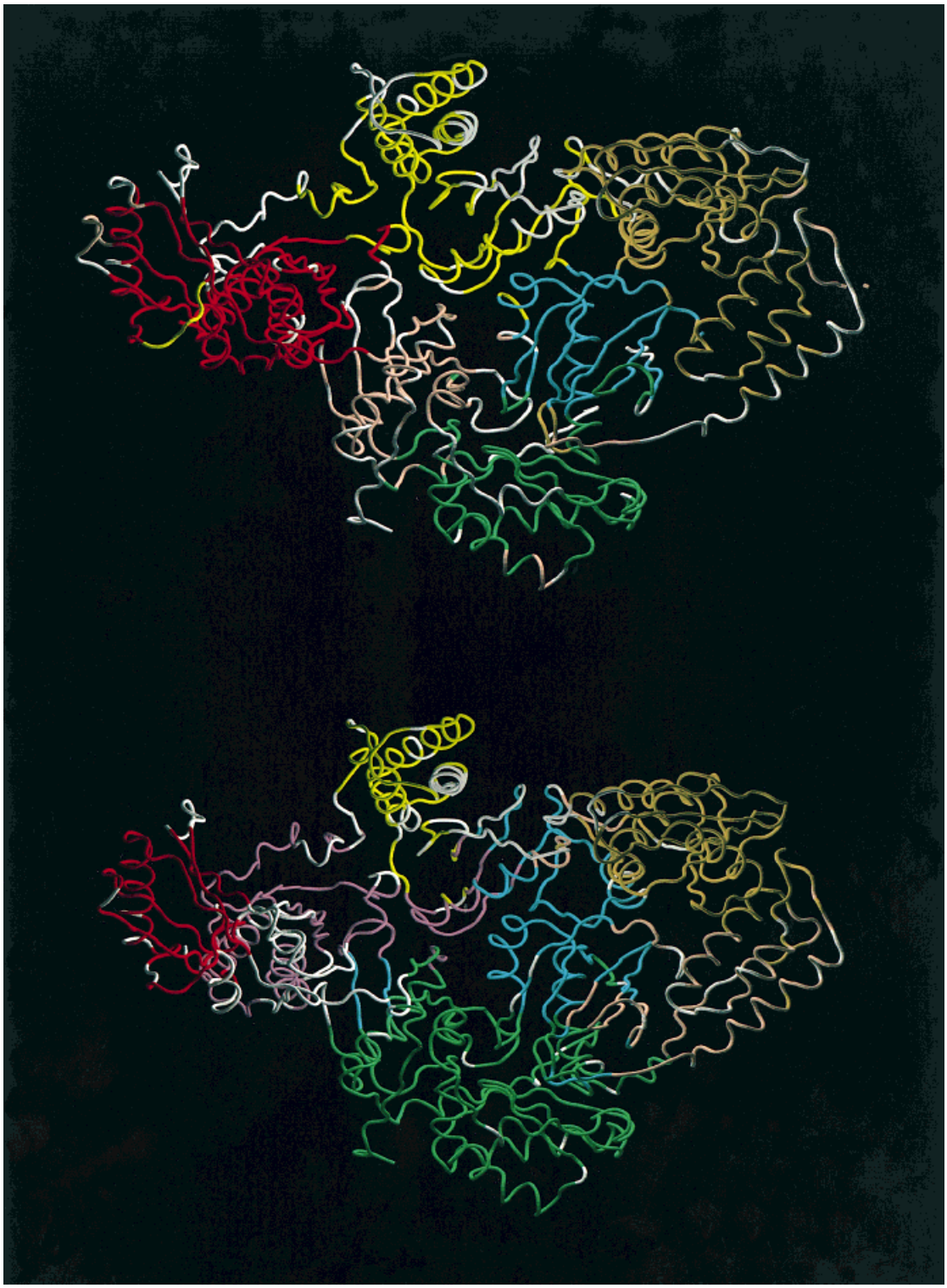

Fig. 4. The domain decompositions of HIV-1 reverse transcriptase, from a comparison of structures with and without a DNA strand (top), and from normal modes (bottom). All colored regions are stable dynamical domains at certain coarseness levels, and the dynamical domains correlate well with the structural domain assignment. The differences between the two domain decompositions can be explained by the different conditions under which the motion was studied, as explained in the Results section, ''HIV-1 reverse transcriptase''.

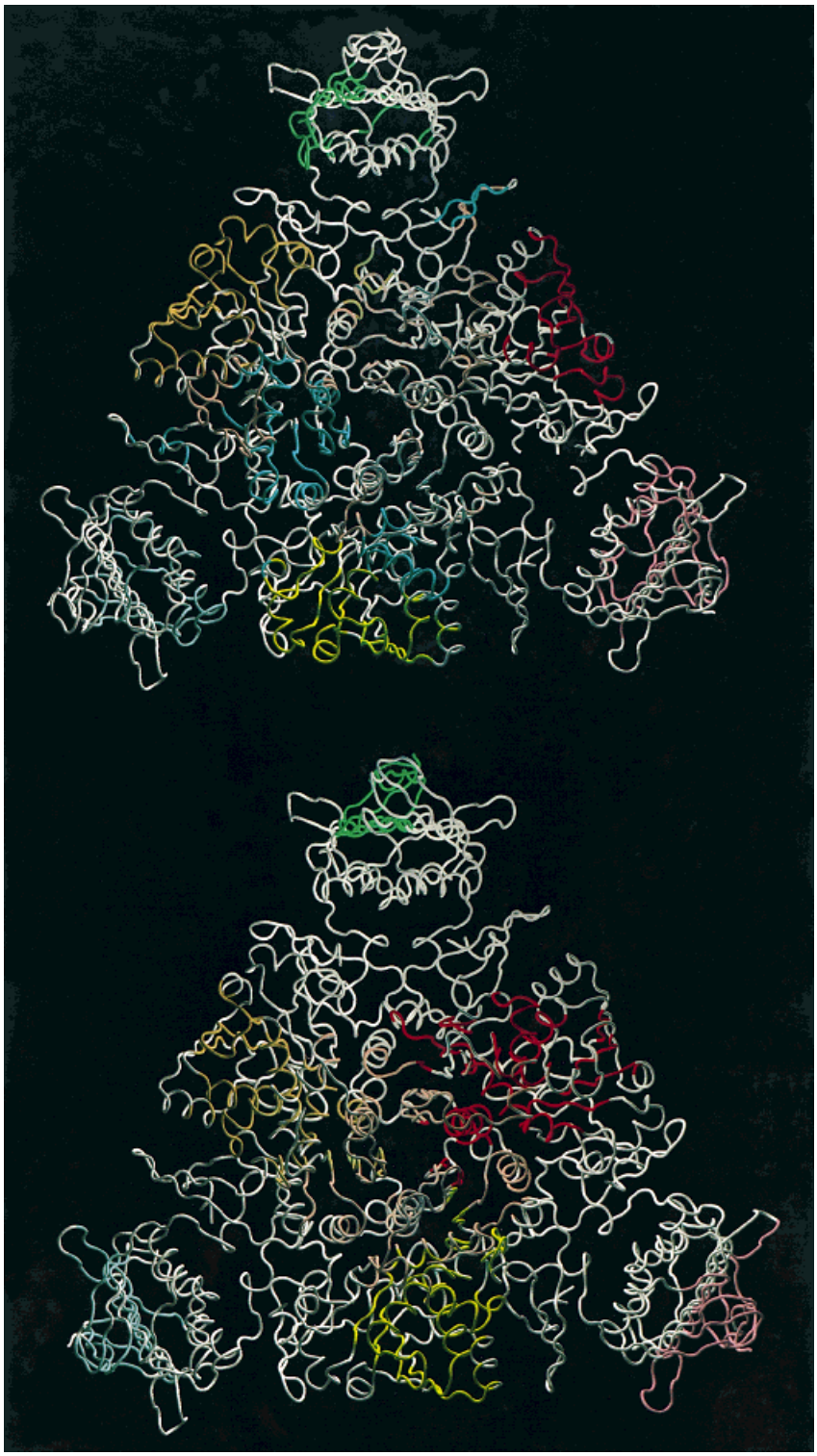

Fig. 5. The domain decompositions of aspartate transcarbamylase, from a comparison of the R and T states (top), and from normal modes (bottom). At higher coarseness values (not shown), each catalytic trimer behaves like one dynamical domain formed by the large and stable interfaces. At lower values, as shown here, the internal structure of the catalytic trimers becomes visible: the central part becomes smaller with decreasing coarseness, whereas the outer parts grow. This indicates that the trimers undergo an overall deformation which accompanies their rigid-body motion. The three regulatory dimers form stable dynamical domains over a large coarseness range.

"fingers," "palm," "thumb," and "connection," and the longer chain p66 has an additional domain "RNAse H," which hydrolyses the RNA strand of the RNA-DNA duplex during first strand synthesis.[23,24]

From the available crystal structures we have chosen PDB entries 1HMI (a complex containing a short DNA strand) and 3HVT (without DNA) for comparison; their C_α-based RMS distance is 0.39 nm. Due to the low resolution of the structure 1HMI, a rather large amount of noise filtering (0.065 nm RMS difference) was necessary to obtain a clear picture. The rigidity threshold was set to 100 kJ/mol. Figure 4 (top) shows the resulting domain decomposition for a coarseness parameter of 10. There is a general similarity to the structural domain decomposition: palm and fingers of chain p66 form one dynamical domain and thumb and connection another one. RNAse H forms a dynamical domain together with the thumb of chain p51. The three remaining structural domains from chain p51 (fingers, palm, and connection) each form a distinct dynamical domain.

A normal mode calculation was performed using the conformation from PDB entry 1HMI and a Fourier basis of 660 modes, corresponding to a cutoff of 1.36 nm. The domain analysis was based on the first seven non-zero modes, using a rigidity threshold of 450 kJ/mol. Since the domain structure of this protein is quite complex, it is interesting to construct a hierarchy of domains, i.e. a description of the breakup of domains at increasing levels of detail. At the coarsest level (coarseness parameter 20), the protein shows only four dynamical domains: (1) the fingers of p66, (2) palm, fingers, and connection of p51, (3) RNAse H and thumb of p51, and (4) palm, thumb, and connection of p66. At a coarseness parameter of 15, the palm of p66 separates from the thumb/connection domain. Reducing the coarseness parameter to 10, we see a separation of thumb and connection of p66, a separation of RNAse H from the p51 thumb, and a separation of the p51 connection domain, which now forms a dynamical domain together with the p66 connection domain. The number of dynamical domains is thus seven at this level, which is shown in Figure 4 (bottom). Decreasing the coarseness parameter further to 5, we see a separation of fingers and palm of p51, and a separation of the two connection domains. We also see the first division within a structural domain: the RNAse H domain splits into two dynamical domains, the smaller one consisting of residues Tyr441 to Thr459, Gly462 to Ala481, and Val536 to Ile556, the larger one consisting of the remaining residues.

Comparing the domain decompositions from the modes and from the conformations, we note that the two are not entirely compatible. The conformation decomposition would have to agree with the mode decomposition at some level of coarseness, but it does not. In the former, the fingers and palm of the p66 chain form one dynamical domain, whereas these two regions belong to different domains in the latter even at a very coarse level. Inversely, fingers and palm of p51 are separated in the conformation comparison, but in the mode analysis they belong to the same domain down to a very low coarseness parameter. However, a perfect agreement is not to be expected in this case, because the two domain decompositions correspond to different physical situations. The normal modes describe the large-scale motions due to thermal agitation, whereas the difference between the two conformations is caused by the presence of a DNA strand. The interactions between protein and DNA are localized to the contact region, whereas the protein-solvent interactions that transmit thermal motion are delocalized. Localized interactions can preferentially cause deformations in places where solvent interactions would show an effect only at higher energies, and vice versa. This explanation is supported by the fact that the domain boundaries that are not identical in the two decompositions are those which are close to the DNA. This example shows that a description of domain motions must always include a specification of the conditions under which the observed motion occurs; there is not necessarily a unique set of low-frequency motions for a given protein.

Aspartate Transcarbamylase

Aspartate transcarbamylase (ATCase) is an allosteric enzyme which catalyses the first step in pyrimidine synthesis. It is a dodecamer consisting of two trimers of catalytic chains and three dimers of regulatory chains. Each chain has two structural domains, named according to the ligand it binds; the regulatory chains have "allosteric" and "zinc" domains, and the domains of the catalytic chains are called "carbamyl phosphate" and "aspartate." The conformations for both the catalytically active R state and the inactive T state are available from crystallography; among the many structures in the PDB, we have chosen the entries 6AT1 (T state without allosteric effector) and 7AT1 (R state complexed with ATP and an analog of the substrate). The allosteric transition in ATCase has been described before;[25,26] the main motions are a relative screw motion of the two catalytic trimers around an axis passing through their centers, consisting of a 1.1 nm translation and a 12 degree rotation, plus a rotation of each regulatory dimer around its axis of pseudo-symmetry by 15 degrees. With 2,778 residues, ATCase is a rather large protein that is difficult to study using conventional numerical approaches such as molecular dynamics or all-atom normal mode analysis.

The two conformations 7AT1 and 6AT1, whose C_α-based RMS distance is 0.63 nm, were filtered up to an RMS difference reduction of 0.03 nm. The rigidity threshold was set to 100 kJ/mol. At high coarseness level, five dynamical domains can be distinguished easily, which correspond to the two catalytic trimers and the outer parts (structurally the allosteric domains) of the three regulatory dimers. The latter remain dynamical domains up to lower coarseness levels, which is not surprising, since an allosteric domain consists essentially of a large 10-stranded beta sheet. The trimers, however, are not so stable; they split into a central region and three outer regions, which grow with decreasing coarseness parameter as the central domain becomes smaller. Figure 5 (top) shows the situation for a coarseness parameter of 7. Here the central domain has also broken up into two parts, and it splits into even more pieces at lower coarseness values. This indicates that the trimers

have an important rigid-body motion, but also an overall deformation distributed more or less evenly. The flexible interface between the trimers and the rigid allosteric domains of the regulatory dimers is formed by the zinc domains of the latter.

The normal modes for ATCase were calculated using the R state conformation 7AT1 and a Fourier basis of 270 basis vectors, corresponding to a cutoff of 2.6 nm. Due to the large size of this protein, many modes must be taken into account to obtain a detailed description of the motion, and consequently a rather high rigidity threshold must be chosen. We used 20 modes and a rigidity threshold of 700 kJ/mol. The domain analysis shows the same behavior as for the conformation comparison: the exterior parts of the regulatory dimers form one stable domain each, whereas each trimer forms one dynamical domain at high coarseness, and a shrinking central part surrounded by growing outer parts with decreasing coarseness. The bottom of Figure 5 shows the situation at a coarseness value of 3. The domain decomposition is also in agreement with an earlier normal mode study using an all-atom model of ATCase.[27]

On the whole, the correspondence between structural and dynamical domains in ATCase is not very pronounced. Although the domain structure of the regulatory chains is clearly visible dynamically, the catalytic trimers behave essentially each like a single dynamical domain. This can be explained by the relatively strong link between the two structural domains of each catalytic chain, formed by two long helices, and by the large and thus strong interface between the three chains that form each trimer.

CONCLUSION

We have presented a collection of techniques that permit a detailed description of low-frequency protein motions in terms of rigid domains, large low-deformation interdomain regions, and highly flexible regions. The analysis can be based either on a comparison of experimentally-obtained structures, or on a simplified normal mode analysis of a single conformation. We have demonstrated the application of this technique to proteins of different size and domain complexity, and shown that conformation comparison and normal modes provide very similar domain decompositions in situations where such a similarity can be expected. Since normal modes provide a clearer and more detailed picture, and since multiple conformations are available experimentally for only a few proteins, a normal-mode-based domain analysis is in general preferable.

The fact that our technique is not computationally expensive, together with the availability of a ready-to-use interactive implementation, means that it can be applied routinely in many situations. In structure determination, it can help to predict whether complexation with a ligand, crystal packing, or other external influences can lead to important conformational changes. In protein engineering, it can indicate whether a given modification is likely to change the dynamical behavior of the protein. In experimental observations of protein motion, it can suggest regions of particular interest. In numerical simulations, it can point out the slow motions whose correct sampling must be verified. And in general, it can help in the process of understanding the relation between protein structure and protein function.

ACKNOWLEDGMENTS

One of the authors (K.H.) thanks the Human Frontier Science Program Organization for a postdoctoral fellowship. We thank the Institut de Biologie Structurale Jean-Pierre Ebel (CEA/CNRS) for support.

REFERENCES

1. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—A Hierarchic Classification of Protein Domain Structures. Structure 1997; 5:1093–1108.
2. Schulz GE. Domain motions in proteins. Curr Opin Struct Biol 1991;1:883–888.
3. Nichols WL, Rose GD, Ten Eyck LF, Zimm BH. Rigid domains in proteins: An algorithmic approach to their identification. Proteins 1995;23:38–48.
4. Wriggers W, Schulten K. Protein domain movements: Detection of rigid domains and visualization of hinges in comparisons of atomic coordinates. Proteins 1997;29:1–14.
5. Hayward S, Kitao A, Berendsen HJC. Model-free methods of analyzing domain motions in proteins from simulations: A comparison of normal mode analysis and molecular dynamics simulation. Proteins 1997;27:425–437.
6. Hayward S, Berendsen HJC. Systematic analysis of domain motions in proteins from conformational change: New results on citrate synthase and T4 lysozyme. Proteins 1998;30:144–154.
7. de Groot BL, Hayward S, van Aalten D, Amadei A, Berendsen HJC. Domain motions in bacteriophage T4 lysozyme: A comparison between molecular dynamics and crystallographic data. Proteins 1998;31:116–127.
8. Gerstein M, Lesk AM, Chothia C. Structural mechanisms for domain movements in proteins. Biochemistry 1994;33:6739–6748.
9. Hinsen K. Analysis of domain motions by approximate normal mode calculations. Proteins 1998;33:417–429..
10. Nishikawa K, Ooi T, Isogai Y, Saito N. Tertiary structure of proteins. I. Representation and computation of the conformations. J Phys Soc Japan 1972;32:1331–1337.
11. Huang ES, Rock EP, Subbiah S. Automatic and accurate method for analysis of proteins that undergo hinge-mediated domain and loop movements Curr Biol 1993;3:740–748.
12. Cornell WD, Cieplak P, Bayly CI, et al., A second generation force field for the simulation of proteins and nucleic acids. J Am Chem Soc 1995;117:5179–5197.
13. Kneller GR. Superposition of molecular structures using quaternions Mol Sim 1991;7:113–119.
14. Jain AK, Dubes RC. Algorithms for clustering data. New York: Prentice-Hall; 1988.
15. Becker OM. Representing protein and peptide structures with parallel coordinates. J Comp Chem 1997;18:1893–1902.
16. Hinsen, K. The Molecular Modeling Toolkit: A case study of a large scientific application in Python. Proceedings of the 6th International Python Conference, http://www.python.org/workshops/1997–10/proceedings/hinsen.html
17. Hinsen K. DomainFinder 1.0, available from http://dirac.cnrs-orleans.fr/DomainFinder
18. Humphrey W, Dalke A, Schulten K. VMD—Visual Molecular Dynamics. J Mol. Graph 1996;14:33–38.
19. Marques O, Sanejouand YH. Hinge bending motion in citrate synthase arising from normal modes calculations. Proteins 1995; 23:557–560.
20. Remington S, Wiegand G, Huber R. Crystallographic refinement and atomic models of two different forms of citrate synthase at 2.7 and 1.7 Å resolution, J Mol Biol 1982;158:111–152.

21. Wiegand G, Remington S. Citrate synthase, structure, control, and mechanism. Annu Rev Biophys Biophys Chem 1986;15:97–117.
22. Bernstein FC, Koetzle TF, Williams GJB, et al. The Protein Data Bank: A computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
23. Wang J, Smerdon SJ, Jäger J, et al. The structural basis of asymmetry in the HIV-1 reverse transcriptase dimer. Proc Natl Acad Sci USA 1994;91:7242–7246.
24. Jäger, J, Smerdon SJ, Wang J, Boisvert DC, Steitz TA. Comparison of three different crystal forms shows HIV-1 reverse transcriptase displays an internal swivel motion. Structure 1994;15:869–876.
25. Lipscomb WN. Aspartate transcarbamylase from Escherichia coli: Activity and regulation. Adv Enzymol Relat Areas Mol Biol 1994;68:67–151.
26. Fetler L, Vachette P, Hervé G, Ladjimi MM. Unlike quaternary structure transition, the tertiary structure change of the 240s loop in allosteric aspartate transcarbamoylase requires active site saturation by substrate for completion. Biochemistry 1995;34: 15654–15660.
27. Thomas A, Field MJ, Mouawad L, Perahia D. Analysis of the low frequency normal modes of the T-state of aspartate transcarbamylase. J Mol Biol 1996;257:1070–1087.