doi:10.1016/S0022-2836(02)00627-7 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) 321, 297–305

JMB

Exploring Global Distortions of Biological Macromolecules and Assemblies from Low-resolution Structural Information and Elastic Network Theory

Florence Tama, Willy Wriggers and Charles L. Brooks III*

Department of Molecular Biology (TPC6), The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA

*Corresponding author

A theory of elastic normal modes is described for the exploration of global distortions of biological structures and their assemblies based upon low-resolution image data. Structural information at low resolution, e.g. from density maps measured by cryogenic electron microscopy (cryo-EM), is used to construct discrete multi-resolution models for the electron density using the techniques of vector quantization. The elastic normal modes computed based on these discretized low-resolution models are found to compare well with the normal modes obtained at atomic resolution. The quality of the normal modes describing global displacements of the molecular system is found to depend on the resolution of the synthetic EM data and the extent of reductionism in the discretized representation. However, models that reproduce the functional rearrangements of our test set of molecules are achieved for realistic values of experimental resolution. Thus large conformational changes as occur during the functioning of biological macromolecules and assemblies can be elucidated directly from low-resolution structural data through the application of elastic normal mode theory and vector quantization.

Keywords: elastic normal modes; vector quantization; codebook vectors; electron microscopy; conformational change

Introduction

In many biological systems, large conformational transitions often involve the relative movement of semi-rigid structural elements. Such motions are important for a variety of protein functions including catalysis and the regulation of activity, as for example in citrate synthase where a hinge motion has been observed upon the binding of coenzyme A.[1,2] These movements are fundamental in the biological function of large and flexible macromolecular complexes, such as motor proteins,[3] chaperonins,[4,5] and the ribosome.[6,7] Thus, the description, prediction and exploration of large-scale conformational distortions of such systems are key in understanding the mechanics of their functioning.

Medium to low-resolution structural information for large macromolecular complexes emerges from a variety of biophysical experiments including X-ray crystallography, electron microscopy and small angle X-ray scattering (SAXS). Large conformational changes in macromolecular complexes are commonly characterized by low-resolution structural methods, in particular by three-dimensional cryogenic electron microscopy (3D cryo-EM).[8] The resolution of 3D cryo-EM reconstruction has been constantly improved over the years: the increasing power of instruments and advanced image processing algorithms now allow one to study very large systems such as viruses at resolutions rivaling X-ray crystallography.[9] Cryo-EM reconstruction is emerging as a primary tool for the structural elucidation of large macromolecular assemblies that are difficult to study by X-ray crystallography and NMR. These techniques are also proving to be powerful in examining the structure and dynamics of macromolecular complexes and their interactions with ligands. One exemplary case comes from the machinery of protein synthesis, where large conformational changes have been observed in the ribosome during the binding of tRNA and protein factors.[10]

Theoretical methods based on atomic or near-atomic theories can be useful in studying these

Abbreviations used: cryo-EM, cryo electron microscopy.

E-mail address of the corresponding author: brooks@scripps.edu

systems and their associated conformational changes. However, conventional molecular dynamics (MD) approaches are presently too costly to be effective for the study of large conformational changes because only a limited range of conformational space can be explored on the timescale of nanoseconds typical for MD studies. An alternative to the direct numerical solution of Newton's equations is the use of normal mode analysis (NMA). This technique has been shown to be very useful in the study of protein motions.[11–13] Even though the harmonic approximation limits the accurate description of energetic landscapes for large conformational displacements, NMA provides information on the preferential direction of collective movements that occur during such displacements. In particular, it has been demonstrated that large conformational changes of proteins observed upon ligand binding,[14–21] and large-scale rearrangements in virus particles[22] can be well represented by the lowest frequency normal modes.

Recent developments in elastic normal mode theory allow calculations on reduced representations of proteins, which include only one point mass per residue,[23] C^{α}-only representations,[24,25] or more coarse-grained particle-based models.[26] In this approach, a simplified potential is used to represent the protein as a set of particles, that describes the mass distribution of the macromolecule, coupled *via* harmonic springs through an elastic "net".[27] One advantage of this model over atomic force field-based normal mode approaches is that it requires no preliminary energy minimization because the "force field" is constructed from the reference configuration and hence is already in its relaxed conformation. Moreover, since the limiting step in NMA is the numeric diagonalization of a $3N$-dimensional matrix, where $3N$ is the number of degrees of freedom, the use of a reduced representation allows one to decrease this complexity. Numerous studies have demonstrated that the motional properties of proteins are reproduced with remarkable fidelity using this simple potential when compared with atomic force field-based models.[23,27] This fact suggests that the essential property one needs to capture in describing global displacements of a protein is the mass distribution of the molecular structure.[24] The extension of such an approach to models of lower resolution may therefore be anticipated to follow from an effective discrete representation of the particle mass distribution.

Many approaches to provide the partitioning of a continuous representation of atomic mass are possible. However, a clustering technique called vector quantization has recently been demonstrated to provide a robust means to develop a discrete reduced representation of continuous 3D data.[28,29] In this approach, the shape of the biological object (molecule, molecular assembly, cellular substructure) is encoded by so-called codebook vectors that identify structural features. As noted above, this description should be sufficient to represent the global distortions of a biological system, since the key information used for reduced representation normal mode calculations in the elastic models is the shape/mass distribution.

Here we describe the development of a framework that combines vector quantization, to yield a discrete reduced description of a continuous shape/mass distribution, with a reduced elastomechanical model for a protein or molecular assembly from which elastic normal modes are calculated to explore the global distortions. We suggest that the synthesis of these two ideas will allow one to predict, explore and rationalize the global distortions and motions of biological structures independent of the resolution of the underlying data set. Low-resolution biophysical data yield valuable information about the architecture of large bio-molecular assemblies, but the motions of such systems have eluded modelers in the absence of a fully atomic detail. We demonstrate for the first time that global distortions of large protein molecules based on calculations utilizing continuous low-resolution data (simulated EM maps) can be captured with high fidelity when compared to motions obtained from more conventional atom-based NMA.

Results

To explore and validate the approach outlined above, proteins that are known to undergo large conformational changes were examined. For these systems, elastic modes were computed for low-resolution data constructed to represent experimental EM density maps and the predicted motions from these representations were compared with those from detailed atomic models of the proteins. These proteins are adenalyte kinase (4ake—214 residues),[30] the maltodextrin binding protein (1omp—370 residues)[31] and citrate synthase (5csc—858 residues).[32]

Normal mode analysis based on atomic-level X-ray structures

NMA was performed on the "open" form of the proteins based on their known X-ray structures. The atom-based normal modes were used as a reference to examine the quality of the global distortions obtained from our discretized elastic model arising from a continuous synthetic low-resolution EM map.

Large conformational changes occur upon ligand binding in each system. The overlap between the vector describing the conformational change, constructed from the difference between the superimposed experimentally determined structures representing the endpoint functional states, and each of the normal modes of the protein was computed. This overlap is a measure of the similarity between the conformational change and the

Table 1. Overlap between the functional conformational changes in 4ake, 1omp, 5csc and each of their lowest-frequency normal modes computed at atomic resolution

Mode	Adenylate kinase	Maltodextrin binding protein	Citrate synthase
1	0.79	0.82	0.01
2	0.32	0.42	0.07
3	0.13	0.05	0.84
4	0.13	0.06	0.01
5	0.30	0.07	0.05

The conformation for adenylate kinase is 1ake, 1anf is for the maltodextrin binding protein and 6csc is for citrate synthase.

structural displacement represented by each of the normal modes. In Table 1, the overlap for the five lowest frequency normal modes is given for each of the proteins. As has been observed,[19,24,33] the overlap between the conformational changes correlates well with one of the low frequency normal modes. These overlaps serve as references for comparison with overlap values obtained from the elastic normal modes based on synthetic EM densities.

Elastic normal modes based on low-resolution structural data

In Figure 1 the synthetic EM map of citrate synthase represented at 15 Å resolution is shown together with two discrete representations of this system, based on the codebook vectors arising from vector quantization of this continuous EM density (see Methods). The elastic models used for the NMA were constructed based on these discrete representations. The overall distribution of the synthetic EM density map is well described by the codebook vectors, even when only 50 vectors are considered (Figure 1(c)).[34]

Since the largest conformational changes are known to be associated with the lowest frequency normal modes, a comparison between the first 20 normal modes obtained from the atomic structure of the protein and for the reduced codebook vector models was made. To examine the agreement between two sets of normal modes, the projection of normal mode i, from the discretized low-resolution representation, onto normal mode j, from the X-ray structure of the protein, is computed, $P_j(i)$. Furthermore, since the modes in one representation may be ordered slightly differently from those in the other, we sum this projection over $2n+1$ modes ($j-n \leq i \leq j+n$; $n=1$ or 2) from the discretized normal mode calculations surrounding mode number j from the atomic representation, i.e. $P_j = \sum_{j-n}^{j+n} P_j(i)$. This projection provides information about how well the normal modes from one representation are expressed in the normal mode basis of the other. It is expected that even if a one-to-one mapping is not observed, the lowest modes from one representation will be well represented by a small number of low modes from the other. The projection over three ($n=1$) and five ($n=2$) modes *versus* the resolution and

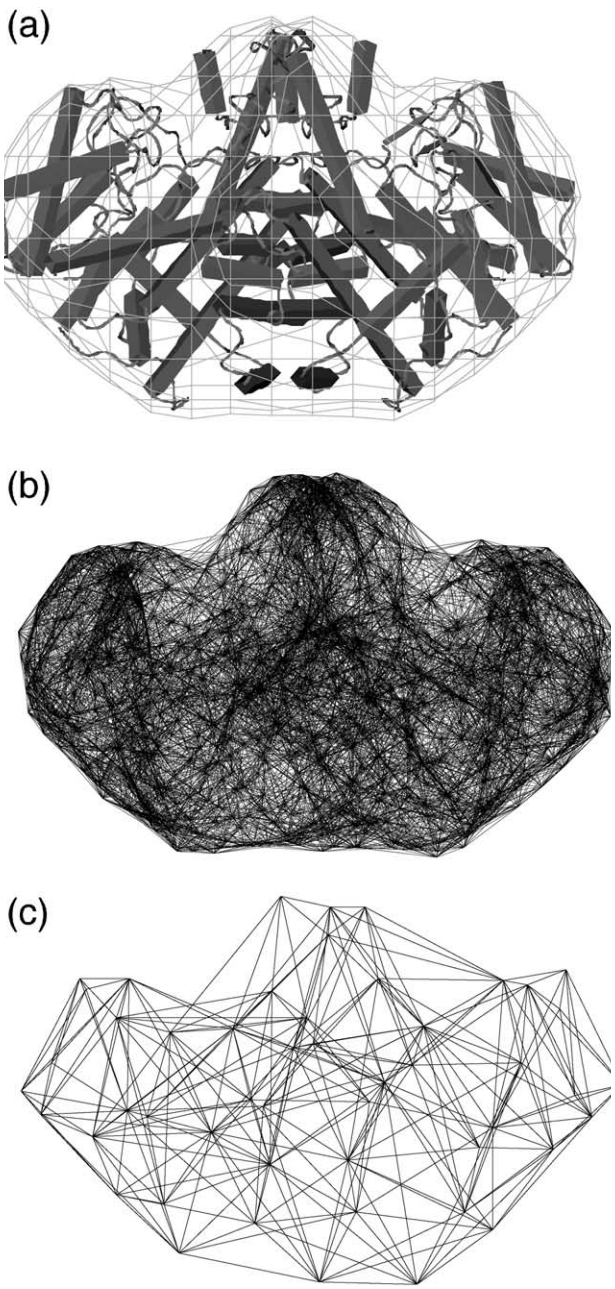


Figure 1. (a) Citrate synthase and its synthetic 3D density at 15 Å resolution. Representative discretized models with (b) 800 codebook vectors and (c) 50 codebook vectors as used for normal mode analysis. The program suite Situs[28] was used for the quantization of volumetric data and the graphics were produced using VMD.[37]

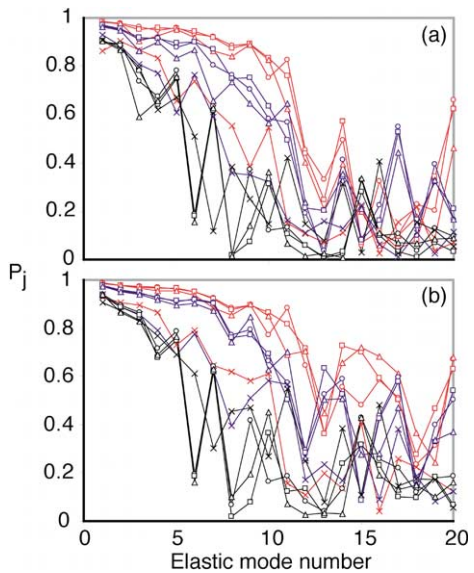


Figure 2. Projection of normal modes from the discretized codebook-vector representation onto normal modes from the atomic model of citrate synthase. The projection for (a) $n = 1$ (see equation (4)) and (b) $n = 2$ at 15 Å resolution (red line), 20 Å resolution (blue line) and 30 Å resolution (black line) using 800 (○), 600 (□), 300 (Δ) and 50 (·) codebook vectors.

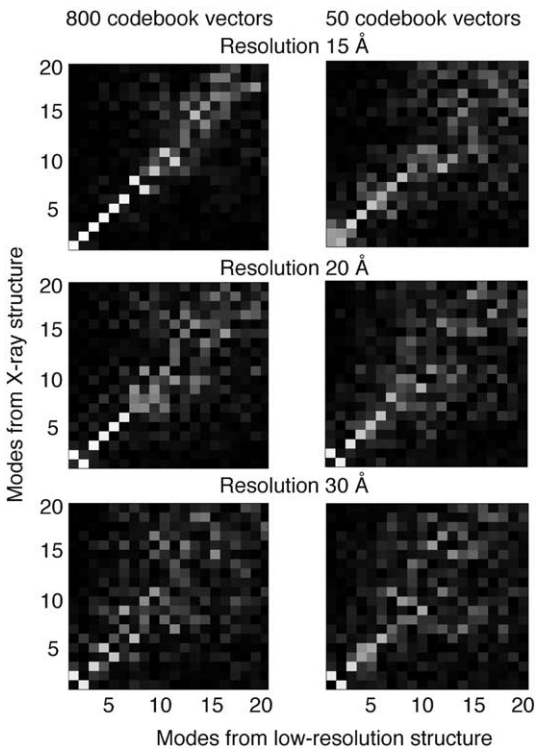


Figure 3. The overlap matrix ($P_j(i)$, see equation (4)) between the first 20 normal modes based on the atomic structure and those from the synthetic low-resolution data of the citrate synthase at 15, 20 and 30 Å resolution for systems with 800 and 50 codebook vectors. White squares correspond to high overlap values (greater than 0.9) and black squares correspond to a low value of the overlap. The grey-scale moves from white to black in ten increments. Optimal agreement between the two sets of modes is observed when the highest overlap is located along the diagonal.

discritization level of the reduced model is shown for citrate synthase in Figure 2. At 15 Å and 20 Å resolution, for systems with 800 or 600 codebook vectors, the projection is almost one for each of the first ten modes. This indicates that the three (or five) modes around mode j in the discretized low-resolution representation provide an excellent description of normal mode i in the atomic representation.

Another means of illustrating the fidelity of the modes *versus* the resolution and complexity of the discretized representation is to display the matrix of overlaps between the modes. In Figure 3(a) the overlap matrices between normal modes from the X-ray structure of the protein and the normal modes obtained from the elastic theory with 800 and 50 codebook vectors at varying resolution are shown. For the model constructed using the highest resolution synthetic EM map, 15 Å, and with the greatest number of codebook vectors (800), the matrix is almost diagonal for the first several modes. In particular, the first six modes from the elastic model have an overlap greater than 0.9 (white squares) with one of the first six low-frequency modes of the atomic structure. This indicates that there is a direct correspondence between the normal modes based on the atomic structure and those obtained from the EM density.

At lower resolution, 20 Å, the quality of agreement for the 800 codebook-vector representation, as measured by a one-to-one mapping, diminishes. This is evident in Figure 3(b) by an increase of the off-diagonal regions of the overlap matrix for the higher modes. However, the first six modes are still in good agreement. At 30 Å resolution, the one-to-one correspondence does not exist and exchange between modes is prevalent. Consistent with the results displayed in Figure 2, the overall description of the normal modes is less precise than at higher resolution.

Describing the system with 300 codebook vectors also gives a good description of the low frequency normal modes (1–10) as seen by a high value P_j (Figure 2). However, the description of higher modes is less precise, and mixing among the modes is more significant, as illustrated by the increase of the value of the projection when considering three overlapping modes *versus* five.

When the representation of the 3D data is reduced to 50 codebook vectors for 15 Å and 20 Å resolution, the elastic model for the normal modes is clearly less precise than with a more complex representation. The off-diagonal parts of the overlap matrix for the first modes increases and the one-to-one correspondence that was observed for 800 codebook vectors all but disappears (Figure 3). At this level of discretization, information on the nature of the global distortions of the biological

system is lost, as indicated by the lower value of the projection in Figure 2. However, at the lowest resolution (30 Å) for a reduced representation at 50 codebook vectors, the agreement is similar to that obtained with representations at 800, 600 or 300 codebook vectors. We observe the same tendency in the cases of the maltodextrin binding protein and adenylate kinase (results not shown).

Describing global conformational distortions with elastic normal modes

Since the prediction and characterization of functionally relevant conformational changes is the main objective of using NMA here, we examine whether the motions suggested from our discrete elastic normal mode description of the EM density agree with the experimentally characterized conformational changes. In Table 2 we show the two modes possessing the largest overlap with the observed conformational change direction as a function of the inherent resolution of the synthetic EM density and the number of codebook vectors utilized in the discretization of this density. Comparison can be made with the overlaps obtained for the normal modes calculated for the X-ray structures of the proteins (Table 1). In each case, a good description of the conformational change is obtained. A high value of overlap is observed for one of the first few lowest normal modes. Even at low resolution or for a small number of codebook vectors, the description of the conformational change is at nearly the same level as that observed from atomic resolution normal modes. We note that the ability of the low resolution model to reproduce the relevant motions described in the atomic structure depends on whether the open or "closed" conformation is used as the basis for computing the modes, but the qualitative agreement between modes from either is reasonable, and the quantitative agreement with modes from an equivalent atomic structure is as noted above (data not shown). Clearly this behavior would deteriorate if the closed conformation possessed a significantly different shape than the open conformation. In Figure 4 we illustrate graphically the normal mode that, at atomic resolution, provides the best representation of the conformational change direction for adenylate kinase. In the same Figure, deformations from the elastic normal mode with the highest overlap (from Table 2) are also shown for systems with 214 and 50 codebook vectors at 15 Å resolution. The global distortions using the elastic theory obtained from the reduced representations of the simulated EM maps reproduce well the functionally relevant motions for this system.

Table 2. Overlap between the functional conformational changes of each protein and the two normal modes from the discrete low-resolution model most involved

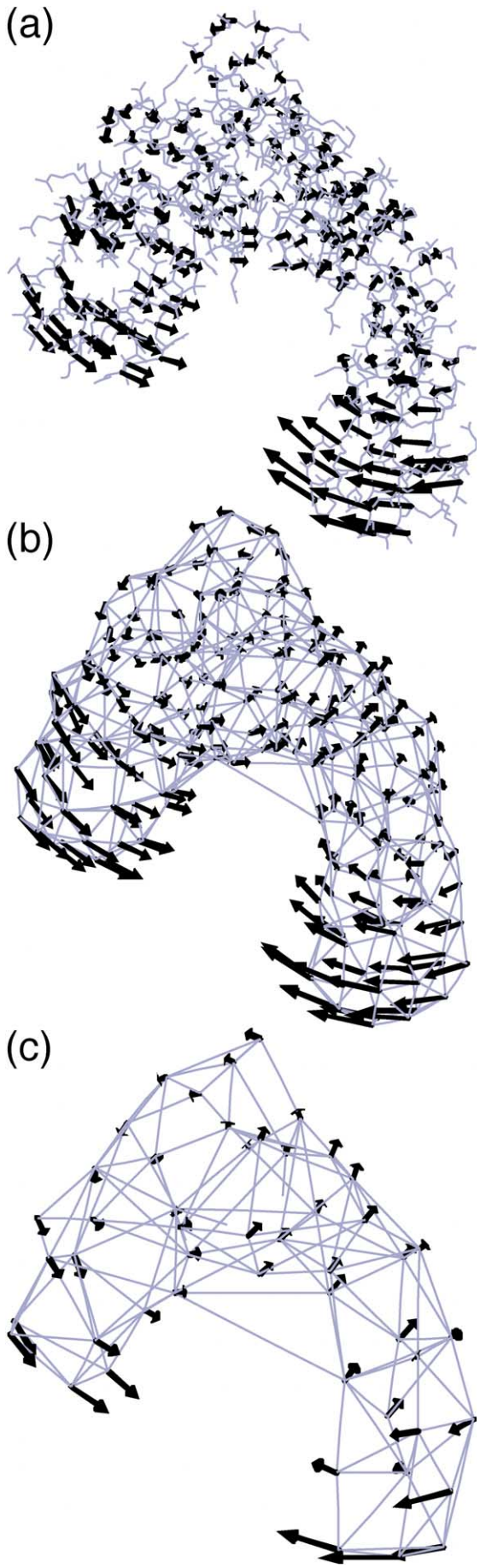
Resolution	15 Å	20 Å	30 Å
Citrate synthase			
800 (12 Å)	0.82 (3)	0.79 (3)	0.73 (3)
	0.12 (2)	0.11 (5)	0.17 (4)
600 (12 Å)	0.82 (3)	0.80 (3)	0.72 (3)
	0.09 (2)	0.12 (1)	0.21 (4)
300 (15 Å)	0.80 (3)	0.79 (3)	0.72 (3)
	0.16 (1)	0.15 (5)	0.20 (4)
50 (25 Å)	0.75 (3)	0.69 (3)	0.70 (3)
	0.29 (1)	0.20 (5)	0.22 (2)
Maltodextrin binding protein			
370 (12 Å)	0.88 (1)	0.82 (1)	0.77 (1)
	0.20 (3)	0.27 (2)	0.31 (2)
200 (12 Å)	0.86 (1)	0.89 (1)	0.78 (2)
	0.26 (3)	0.13 (3)	0.31 (3)
100 (18 Å)	0.81 (1)	0.69 (2)	0.82 (1)
	0.31 (2)	0.51 (1)	0.23 (3)
50 (18 Å)	0.81 (2)	0.77 (1)	0.73 (1)
	0.24 (3)	0.21 (2)	0.37 (2)
Adenylate kinase			
214 (12 Å)	0.78 (1)	0.72 (1)	0.58 (1)
	0.36 (2)	0.42 (2)	0.47 (2)
100 (15 Å)	0.76 (1)	0.70 (1)	0.66 (2)
	0.38 (2)	0.42 (2)	0.36 (1)
50 (18 Å)	0.75 (1)	0.69 (2)	0.57 (1)
	0.40 (2)	0.46 (1)	0.42 (2)

Overlap is shown as a function of the representation and the resolution. The cut-off and mode number are indicated in parentheses.

Discussion

The quality of the agreement between the elastic modes as a function of the underlying resolution of the EM density and the number of codebook vectors is clearly illustrated in Figure 2. For maps at resolutions of 15 Å and 20 Å, remarkably good agreement can be achieved with representations using 800 and 600 codebook vectors. In particular, the first ten lowest modes obtained from the elastic network NMA are highly overlapping with those obtained from the high-resolution structure. For each of the three proteins studied here, significant fidelity in the description of the functionally relevant conformational changes was also found for these models. This agreement is sufficient to suggest a high level of confidence in using such approaches to explore functional motions from structural models lacking atomic detail such as derived from EM. Moreover, we anticipate, based on studies of highly symmetric virus particles using similar methods with atomic models, that good correspondence with global distortions of the system will be attainable for such systems as well.[22] However more severe discretizations, such as the 50-codebook-vector representation for both 15 Å and 20 Å resolution synthetic EM maps, inhibit the reproduction of global distortions of the biological object.

As the resolution of the EM maps is lowered to 30 Å, the one-to-one correspondence between modes from the discretized model and the atomic model disappears and significant mixing among the modes is seen. As indicated in Figure 2, the



overall description of the global distortion of the system is less precise than at higher resolution.

Naturally, high-resolution EM maps are necessary to obtain precise information on functionally relevant conformational changes. They provide more complete descriptions of the overall electron distribution of the system. However, experimental difficulties often inhibit the establishment of high-resolution data. Thus, exploration of medium-resolution data requires more careful consideration of the underlying parameters of the theory.

The choice of the number of codebook vectors will depend on the resolution, shape and size of the molecule or assembly. In the case of citrate synthase, 800 or 600 codebook vectors at 15 Å or 20 Å resolution guarantees a good description of the lowest vibrational modes (up to ten). This protein comprises 858 residues, so 70% of the total number of residues is sufficient to yield significant results. A good description for fewer elastic modes is still observed for 300 codebook vectors, which corresponds to 35% of the total number of residues. In the case of maltodextrin binding protein (results not shown), at 15 Å and 20 Å resolution we observed that the best description occurred when 370 codebook vectors were employed: however, good agreement was still obtained for 200 codebook vectors, which corresponds to 55% of the total number of residues. For adenylate kinase, a 100 codebook-vector representation gave normal modes in quite good agreement with those based on the atomic structure. This corresponds to a reduction by 50%. At lower resolution, 30 Å, the number of codebook vectors does not affect the result, as previously observed for citrate synthase.

The results clearly indicate that models with 50 or fewer codebook vectors are not sufficient for EM maps at moderate resolutions of 15–20 Å. At this resolution, a more detailed description of the density distribution of the biological object is present. Thus, more features in the data can be identified and the number of codebook vectors needed to adequately describe these data is larger. Thus, at higher resolutions the most appropriate choice for the number of codebook vectors, which guarantees good fidelity in the normal modes, corresponds roughly to one point per residue. At lower resolution, the number of features present in the EM maps is smaller and fewer codebook vectors are necessary for an adequate representation of the discretized density.

Within the context of the elastic network model, the number of cluster points (codebook

Figure 4. Amplitude and direction of motion for the normal mode that best overlaps with the conformational change observed experimentally for the adenylate kinase from calculations based on (a) the X-ray structure, (b) a 214 codebook-vector representation, and (c) a 50 codebook-vector representation.

vectors) is linked to an appropriate choice of a cut-off for inter-cluster connections, and hence the connectedness of the elastic net representation. As described in Methods, the choice of an optimal cut-off in establishing this connectedness is related to the distribution of codebook-vector separations. Quite generally, choosing a value for the cut-off that is after the second peak in this distribution will provide a reasonable model. However, for extreme discritizations (less than 50 codebook vectors) artifacts can arise because of the incompleteness of the discrete representation. Thus, even if the number of codebook vectors required to represent the electron density for data at 30 Å resolution is less than the number of residues of the system, it may be wisest to use one point per residue. With one point per residue (or basic unit of mass, e.g. amino acid, nucleic acid base, etc.), a cut-off of 12 Å is sufficient to yield quite reasonable elastic normal mode models. For very large assemblies, this prescription may lead to an inhibitive numerical diagonalization. In such cases, a reduction of the discretized system by up to 50% will still yield a robust description of the global distortions of the biological object while permitting a cut-off in the range of 12–15 Å to be employed. Ideally, one would like to determine the connectivity of adjacent codebook vectors automatically based on geometric considerations,[34] but this formulation would depart from the Tirion model employed here and is left as the subject of future research.

Conclusions

NMA on low-resolution structural data to explore the global, and functionally relevant, distortions of large biological molecules and assemblies can be successfully performed in the absence of detailed atomic models by combining the methods of discretized representations provided by vector quantization[28] with elastic network theories.[27] The present study provides the first demonstration of this. Studies of conformational changes in biological systems using normal mode theory need no longer be limited by the absence of high-resolution X-ray crystallographic structures. The approach we describe here opens the door to further studies aimed at understanding the mechanisms of action in large assemblies such as the ribosome, for which experimental data from cryo-EM is available for different functional (and conformational) states.

Methods

Elastic network normal mode analysis

NMA is a common tool to study protein dynamics and more particularly large conformational changes.[24] Recently, a simplified representation of the potential energy has been used in normal mode calculations on biological molecules at atomic resolution.[27] In this representation the protein is described as a three-dimensional elastic network based on the equilibrium distribution of atoms.

In the elastic network model, amino acids may be represented in full atomic detail or reduced to a single coordinate (one point mass per residue or the C^{α} atom positions).[23–25] The positions of these sites identify the junctions within the network. Coarser grained models, i.e. where between $N/2$ and $N/40$ atoms are used to identify the junctions of the network, have also been considered.[26] These junctions are representative of the mass distribution of the system and are linked together by harmonic springs using a Hookean pairwise potential:

$$U(\mathbf{r}_a, \mathbf{r}_b) = \frac{C}{2}(|\mathbf{r}_{a,b}| - |\mathbf{r}^{\mathbf{0}}_{a,b}|)^2 \qquad (1)$$

where $\mathbf{r}_{a,b} = \mathbf{r}_a - \mathbf{r}_b$ denotes the vector connecting two junctions, a and b, and the zero superscript indicates the given initial configuration. The strength of the potential, C, is a phenomenological constant assumed to be the same for all interacting junctions, and is set to 1 in the current calculations. Within this description, the total potential energy of the system is given by:

$$U^{\text{total}} = \frac{1}{2}\sum_{a,b} U(\mathbf{r}_a, \mathbf{r}_b)\theta(R_{\text{cut-off}} - |\mathbf{r}^{\mathbf{0}}_{a,b}|) \qquad (2)$$

The sum is restricted to pairs separated by less than $R_{\text{cut-off}}$, which is a parameter describing the effective interaction length-scale, by the step function $\theta(x)$ The step-function adopts a value of one when its argument is greater than zero and is zero everywhere else.

Vector quantization

The potential energy function for an atomically represented mass distribution introduced above can be transformed to one capable of representing a low-resolution continuous distribution of density. To do this one can work either in a continuum representation, or transform the continuum representation into a discrete one. We pursue the latter idea and associate the junctions (mass points) of the network with the codebook vectors obtained from vector quantization.[28] Vector quantization applied to 3D data, such as the continuous density representing molecular structures from the experimental technique of cryo-EM, provides a discrete reduced representation that is suitable for the development of low-resolution models.

Normal mode analysis: X-ray structure

NMA was performed using a combination of the elastic network model, where each junction is identified with a heavy atom of the protein, and the rotation–translation block (RTB) method, using one block per residue.[33,35] A cut-off, $R_{\text{cut-off}}$, of 8 Å was used in the calculations on atomically detailed models.

Elastic network theory: low-resolution data

For each protein, low-pass filtered synthetic electron density maps at 15 Å, 20 Å and 30 Å resolution were created with the pdblur utility implemented in the Situs package.[28,29] The voxel spacing for the maps was set to 2 Å. The atomic structure was convolved with a Gaussian kernel of variable width. The density cut-off

of the synthetic maps was chosen in a way that each system exhibited a resolution-independent volume. To obtain a reduced representation of an electron density map, a vector quantization was performed with the qvol utility implemented in the Situs package. Different numbers of codebook vectors were placed at the features of a given 3D density distribution (vector quantization).[28] The codebook vectors form a set of control points or landmarks that provide information about the shape and the density distribution of a biological object. For each synthetic map, data sets were generated with varying numbers of codebook vectors utilizing the open forms of each protein: for adenylate kinase, three data sets were generated with 214, 100 and 50 codebook vectors; for the maltodextrin binding protein 370, 200, 100 and 50 codebook vectors were used; discretized models with 800, 600, 300 and 50 codebook vectors were generated for citrate synthase.

For each representation, different values of the distance cut-off, $R_{\text{cut-off}}$, were used to determine when pairs of codebook vectors are supposed to be linked together by a harmonic spring. $R_{\text{cut-off}}$ is an important parameter in the elastic NMA, since it determines the number of interactions (or links) between each codebook vector. Generally, its value should be chosen to be after the second peak in the distribution of codebook-vector center–codebook-vector center separations for the model (or atom–atom separations in atomic models). It has been shown that when only C^{α} atoms are used, the best results are obtained with a cut-off of 12–13 Å, which corresponds to the point in the C^{α}–C^{α} distribution of separations just noted. In our calculations using simulated EM densities, some models had sparsely distributed codebook vectors and the distance between each point in these models was significantly larger than the mean C^{α}–C^{α} distance in proteins. For models with few codebook vectors, a 12–13 Å cut-off is not appropriate and can give rise to artifacts. Thus different cut-off values ranging from 12 Å to 30 Å were used depending on the number of codebook vectors and corresponding codebook-vector separation distribution in the system. When the number of codebook vectors is similar to the number of residues in the protein, we observed almost no difference in the normal modes using a cut-off of 12 Å, 15 Å, or 18 Å. Decreasing the number of codebook vectors to 300, required a minimum cut-off of 15 Å to avoid artifacts arising from too sparse a connectivity. Similar agreement was obtained with a cut-off of 18 Å. For systems with only 50 codebook vectors, we observed for all resolutions (15 Å, 20 Å and 30 Å) that a cut-off value of 22 Å was insufficient to yield a good description of normal modes. For example, in the case of the citrate synthase at 30 Å resolution, the maximum projection (P_j) obtained with the 22 Å cut-off was 0.45. The best agreement occurred when the cut-off was increased to 25 Å. For cut-off values significantly greater than 25 Å, poorer quality agreement was observed because the longer cut-off lead to too great a connectivity in the elastic network. For these extreme discritizations of the low-resolution EM maps, even though we found a significant variation in the quantitative behavior of the normal modes, the qualitative character of the displacements were not nearly as sensitive to the particular value chosen for the cut-off.

Normal modes obtained from the low-resolution representation yield a displacement vector for each of the codebook vectors in the system. In order to compare these elastic distortions with normal modes obtained directly from the atomic structure, an interpolation of the motion of the codebook vectors to the motion of the atoms is needed. An interpolation based on 3D thin plate splines[36] was used to extend the sparsely sampled displacement vector fields to the atoms of the proteins (Chacon & Wriggers, unpublished results). The displacements thus generated from the interpolation method are compared with the displacements observed from NMA for the X-ray structure.

Analysis

To quantify how a given normal mode, $\mathbf{a}_j$, compares with an experimentally known conformational change between an open and closed conformation, $\Delta\mathbf{r} = \mathbf{r}_o - \mathbf{r}_c$, the overlap between the two corresponding vectors can be calculated as:

$$I_j = |\mathbf{a}_j \cdot \Delta\mathbf{r}| = \frac{\left|\sum \alpha_j^i \cdot (\mathbf{r}_i^o - \mathbf{r}_i^c)\right|}{\sqrt{\sum_j \sum_i \left|\alpha_j^i \cdot (\mathbf{r}_i^o - \mathbf{r}_i^c)\right|^2}} \qquad (3)$$

where $\mathbf{r}_i^o$ and $\mathbf{r}_i^c$ are, respectively, the coordinates of protein atom i in conformations o and c, after superposition, and α_j^i is the coefficient of the jth normal mode along the ith atomic direction. An overlap of one would indicate that a given mode perfectly captures the collective atomic displacements.

The normal mode vector from the atomic representation, $\mathbf{a}_j$, can be expressed as a linear combination of $2n + 1$ normal modes $\mathbf{b}_i$ obtained from the low-resolution discretized representation of the protein. If the displacements represented in each of the different normal mode models are similar, the sum of this "projection" of one space onto the other will approach a value that is close to one. We define P_j as a measure of this overlap by:

$$P_j = \sum_{i=j-n}^{j+n} P_j(i) = \sum_{i=j-n}^{j+n} (\mathbf{a}_j \cdot \mathbf{b}_i)^2 \qquad (4)$$

It is P_j that is plotted in Figure 2 *versus* mode number j and for $n = 1, 2$. $P_j(i)$ is represented by a given matrix element plotted in Figure 3.

Acknowledgments

The authors thank Pablo Chacon for helpful discussions. Financial support from the National Institutes of Health (RR12255 to C.B. and W.W., and GM62968 to W.W.) is appreciated.

References

1. Remington, S., Wiegand, G. & Huber, R. (1982). Crystallographic refinement and atomic models of 2 different forms of citrate synthase at 2.7 Å and 1.7 Å resolution. *J. Mol. Biol.* **158**, 111–152.
2. Huber, R. & Bennett, W. S. (1983). Functional-significance of flexibility in proteins. *Biopolymers*, **22**, 261–279.
3. Jontes, J. D. & Milligan, R. A. (1997). Brush border myosin-I structure and ADP-dependent conformational changes revealed by cryoelectron microscopy and image analysis. *J. Cell Biol.* **139**, 683–693.
4. Roseman, A. M., Chen, S. X., White, H., Braig, K. & Saibil, H. R. (1996). The chaperonin ATPase cycle:

mechanism of allosteric switching and movements of substrate-binding domains in GroEL. *Cell*, **87**, 241–251.
5. Xu, Z. H., Horwich, A. L. & Sigler, P. B. (1997). The crystal structure of the asymmetric GroEL-GroES-(ADP)(7) chaperonin complex. *Nature*, **388**, 741–750.
6. Frank, J. & Agrawal, R. K. (2000). A ratchet-like inter-subunit reorganization of the ribosome during translocation. *Nature*, **406**, 318–322.
7. Agrawal, R. K., Heagle, A. B., Penczek, P., Grassucci, R. A. & Frank, J. (1999). EF-G-dependent hydrolysis induces translocation accompanied by large conformational changes in the 70 S ribosome. *Nature Struct. Biol.* **6**, 643–647.
8. Saibil, H. R. (2000). Conformational changes studied by cryo-electron microscopy. *Nature Struct. Biol.* **7**, 711–714.
9. Chiu, W., McGough, A., Sherman, M. B. & Schmid, M. F. (1999). High-resolution electron cryomicroscopy of macromolecular assemblies. *Trends Cell Biol.* **9**, 154–159.
10. Frank, J. (2001). Cryo-electron microscopy as an investigative tool: the ribosome as an example. *Bioessays*, **23**, 725–732.
11. Go, N., Noguti, T. & Nishikawa, T. (1983). Dynamics of a small globular proteins in terms of low-frequency vibrational modes. *Proc. Natl Acad. Sci. USA*, **80**, 3696–3700.
12. Brooks, B. R. & Karplus, M. (1983). Harmonic dynamics of proteins: normal mode and fluctuations in bovine pancreatic trypsin inhibitor. *Proc. Natl Acad. Sci. USA*, **80**, 6571–6575.
13. Levitt, M., Sander, C. & Stern, P. S. (1985). Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *J. Mol. Biol.* **181**, 423–447.
14. Harrison, W. (1984). Variational claculation of the normal modes of a large macromolecule: methods and some initial results. *Biopolymers*, **23**, 2943–2949.
15. Brooks, B. R. & Karplus, M. (1985). Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme. *Proc. Natl Acad. Sci. USA*, **82**, 4995–4999.
16. Gibrat, J. F. & Go, N. (1990). Normal mode analysis of human lysozyme: study of the relative motion of the two domains and characterization of the harmonic motion. *Proteins: Struct. Funct. Genet.* **8**, 258–279.
17. Seno, Y. & Go, N. (1990). Deoxymyoglobin studied by the conformational normal mode analysis. 1. Dynamics of globin and the heme–globin interaction. *J. Mol. Biol.* **216**, 95–109.
18. Seno, Y. & Go, N. (1990). Deoxymyoglobin studied by the conformational normal mode analysis. 2. The conformational change upon oxygenation. *J. Mol. Biol.* **216**, 111–126.
19. Marques, O. & Sanejouand, Y. H. (1995). Hinge-bending motion in citrate synthase arising from normal mode calculations. *Proteins*, **23**, 557–560.
20. Perahia, D. & Mouawad, L. (1995). Computation of low-frequency normal-modes in macromolecules—improvements to the method of diagonalization in a mixed basis and application to hemoglobin. *Comput. Chem.* **19**, 241–246.
21. Mouawad, L. & Perahia, D. (1996). Motions in hemoglobin studied by normal mode analysis and energy minimization: evidence for the existence of tertiary T-like, quaternary R-like intermediate structures. *J. Mol. Biol.* **258**, 393–410.
22. Tama, F., Brooks, C. L. & II, . (2002). The mechanism and pathway of pH induced swelling in cowpea chlorotic mottle virus. *J. Mol. Biol.* **318**, 733–747.
23. Hinsen, K. (1998). Analysis of domain motions by approximate normal mode calculations. *Proteins: Struct. Funct. Genet.* **33**, 417–429.
24. Tama, F. & Sanejouand, Y. H. (2001). Conformational change of proteins arising from normal mode calculations. *Protein Eng.* **14**, 1–6.
25. Bahar, I., Atilgan, A. R. & Erman, B. (1997). Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.* **2**, 173–181.
26. Doruker, P., Jernigan, R. L. & Bahar, I. (2002). Dynamics of large proteins through hierarchical levels of coarse-grained structures. *J. Comput. Chem.* **23**, 119–127.
27. Tirion, M. M. (1996). Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Letters*, **77**, 1905–1908.
28. Wriggers, W. & Birmanns, S. (2001). Using Situs for flexible and rigid-body fitting of multiresolution single-molecule data. *J. Struct. Biol.* **133**, 193–202.
29. Wriggers, W., Milligan, R. A. & McCammon, J. A. (1999). Situs: A package for docking crystal structures into low-resolution maps from electron microscopy. *J. Struct. Biol.* **125**, 185–195.
30. Muller, C. W., Schlauderer, G. J., Reinstein, J. & Schulz, G. E. (1996). Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure*, **4**, 147–156.
31. Sharff, A. J., Rodseth, L. E., Spurlino, J. C. & Quiocho, F. A. (1992). Crystallographic evidence of a large ligand-induced hinge-twist motion between the two domains of the maltodextrin binding protein involved in active transport and chemotaxis. *Biochemistry*, **31**, 10657–10663.
32. Liao, D. I., Karpusas, M. & Remington, S. J. (1991). Crystal structure of an open conformation of citrate synthase from chicken heart at 2.8 Å resolution. *Biochemistry*, **30**, 6031–6036.
33. Tama, F., Gadea, F. X., Marques, O. & Sanejouand, Y. H. (2000). Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins: Struct. Funct. Genet.* **41**, 1–7.
34. Wriggers, W., Milligan, R. A., Schulten, K. & McCammon, J. A. (1998). Self-organizing neural networks bridge the biomolecular resolution gap. *J. Mol. Biol.* **284**, 1247–1254.
35. Durand, P., Trinquier, G. & Sanejouand, Y. H. (1994). New approach for determining low-frequency normal-modes in macromolecules. *Biopolymers*, **34**, 759–771.
36. Bookstein, F. L. (1989). Principal warps: thin-plate splines and the decompsosition of deformations. *IEEE Trans. Patt. Anal. Mach. Intell.* **11**, 567–586.
37. Humphrey, W., Dalke, A. & Schulten, K. (1996). VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38.

Edited by W. Baumeister

(Received 23 April 2002; received in revised form 14 June 2002; accepted 17 June 2002)